

Charles University
Faculty of Science

Study programme: Philosophy and History of Science

Branch of study: Philosophy and History of Science



Mgr. Michaela Jirout Košová

Folk Dualism and the Two Conceptual Realms
Lidový dualismus a dvě konceptuální říše

Doctoral thesis

Supervisor: prof. RNDr. Jaroslav Peregrin, CSc.

Prague, 2020

Prohlášení

Prohlašuji, že jsem disertační práci vypracovala samostatně, že jsem všechny použité zdroje a literaturu řádně citovala a že tato práce ani její podstatná část nebyla využita jako závěrečná práce k získání jiného nebo stejného akademického titulu.

V Praze dne 31. 8. 2020

Mgr. Michaela Jirout Košová

Acknowledgement

I would like to thank my supervisor prof. Jaroslav Peregrin for introducing me to ideas that became the cornerstone of this thesis and providing me with thoughtful comments and suggestions at all stages of writing. I would also like to thank the Charles University Grant Agency and Hlávka Foundation for enabling me to meet and share thoughts with the best researchers in experimental philosophy from all around the world. Many thanks to my colleagues from the department for their support and inspiration during our seminars. A special thanks goes to my colleague and good friend Robin Kopecký for encouraging me to become an experimental philosopher, for his trust, support, and countless inspiring ideas throughout all these years of our friendship and collaboration on research projects. Last but not least, many thanks to my husband Vašek and the rest of my loved ones for their trust, support, and patience during all those long hours and days I spent with this thesis instead of with them.

Abstract

The thesis focuses on the irreducibility of the concept of a person to scientific view of the world. The main inspiration for thematising this specific aspect of folk dualism comes from Donald Davidson (*two realms*) and Wilfrid Sellars (*two images*). The theoretical sections are complemented by reflexion on results of empirical studies provided mostly by experimental philosophy in order to demonstrate how this approach benefits attempts to reach complex view of philosophical questions that have close connection to moral dimension of human life. The first chapter addresses a wider concept of self and introduces the idea of the necessity to bring the two conceptual realms on the scene: there is a specific conceptual realm (irreducible to physical realm or scientific image) enabling proper grasp of the concept of a person. The subsequent chapters address particular sub-concepts of the concept of self. The second chapter focuses on the concept of free will, and by referring to different views it points to the necessity to bring folk concepts into consideration. It concludes that the folk concept of free agent is transcendent with regard to scientific accounts and bears certain “supernatural” characteristics connected to the concept of conscious will. The third (and central) chapter brings focus on the concept of the essential self. Via reflecting on the research in the area of folk dualism and folk intuitions about personal identity it uncovers the normative core of the folk concept of a person, namely the moral and interpersonal dimension that comes forward in the concept of *the essential moral self* and in the folk concept of soul. The last chapter deals with the concept of consciousness. By addressing different empirical studies, it criticises and revises the role that the philosophically understood problem of consciousness plays in the folk concept of a person. The main conclusion of this chapter is that the folk don’t seem to acknowledge the hard problem of consciousness in the same way as philosophers do. Instead, only certain specific conscious states are the key to the folk concept of a person, thanks to their connection to moral and interpersonal dimension. The conclusion of the thesis brings forward the importance of exploring folk intuitions and distinguishing between the two conceptual realms in solving moral questions not only for today, but also for the future.

Keywords: Donald Davidson, consciousness, experimental philosophy, folk concepts, folk dualism, free will, person, personal identity, self, soul, Wilfrid Sellars

Abstrakt

Tato práce je zaměřená na neredukovatelnost konceptu osoby na vědecký popis světa. Hlavní inspirace pro tematizaci tohoto specifického aspektu lidového dualismu přichází od Donalda Davidsona (*dvě říše*) a Wilfrida Sellarse (*dva obrazy*). Teoretické pasáže jsou doplněny reflexí nad výsledky empirických studií vycházejících zejména z experimentální filozofie s cílem demonstrovat přínos tohoto přístupu v pokusech dosáhnout komplexního náhledu na filosofické otázky spojené s morální dimenzí lidského života. První kapitola pojednává o širším konceptu Já a poukazuje na nutnost uvést na scénu dvě konceptuální říše: existuje specifická konceptuální říše (neredukovatelná na fyzikální říši a vědecký obraz), která umožňuje adekvátní uchopení konceptu osoby. Následující kapitoly rozebírají jednotlivé subkoncepty konceptu Já. Druhá kapitola je zaměřená na koncept svobody vůle a skrze probírání různých přístupů poukazuje na nutnost vzít do úvahy lidové koncepty. Dospívá k závěru, že lidový koncept svobodného konatele je s ohledem k vědeckým přístupům transcendentní a vykazuje jisté „nadpřirozené“ charakteristiky spojené s konceptem vědomé vůle. Třetí (a centrální) kapitola je zaměřená na koncept esenciálního Já. Reflexí nad empirickými výzkumy lidového dualismu a lidových intuicí týkajících se osobní identity odhaluje normativní jádro lidového konceptu osoby, a sice morální a mezilidskou dimenzi vynořující se z konceptu *esenciálního morálního Já* a z lidového konceptu duše. Poslední kapitola pojednává o konceptu vědomí. Odkazujíc se na různé empirické studie kritizuje a klade na pravou míru roli, kterou filozoficky chápaný problém vědomí hraje v rámci lidového konceptu osoby. Hlavním závěrem kapitoly je tvrzení, že nefilozofové neuznávají těžký problém vědomí do takové míry, jak je tomu u filozofů. Jenom některé specifické vědomé stavy jsou klíčem k lidovému konceptu osoby, a to díky jejich spojení s morální a mezilidskou dimenzí. Závěr práce vyzdvihuje důležitost zkoumání lidových intuicí a rozlišování mezi dvěma konceptuálními říšemi v řešení morálních otázek nejenom pro dnešek, ale i do budoucnosti.

Klíčová slova: Donald Davidson, duše, experimentální filozofie, Já, lidové koncepty, lidový dualismus, osoba, osobní identita, svoboda vůle, vědomí, Wilfrid Sellars

Contents

Contents	11
Introduction.....	13
Two realms, two images	15
Experimental philosophy	20
1 The Concept of Self.....	27
1.1 Transcendental naturalism	28
1.2 Self as a useful construction.....	33
1.3 Anomalous monism	40
1.4 Self and the two conceptual realms.....	44
2 The Concept of Free Will.....	48
2.1 Apparent compatibilism	50
2.2 Changing the subject.....	57
2.3 Free will and the two conceptual realms.....	60
2.4 Support from experimental philosophy.....	66
3 The Concept of Soul and the Essential Moral Self.....	74
3.1 The complexity of folk dualism: Mind vs. soul	75
3.2 Personal identity and the essential moral self	80
3.3 The essential moral self and folk dualism: Experimental studies	90
4 The Concept of Consciousness.....	102
4.1 Consciousness-brain dualism and the explanatory gap.....	103
4.2 Experimental philosophy of consciousness	109
4.3 How philosophers and the folk see consciousness.....	124
Conclusion	136
Bibliography	143

Introduction

“Hence it is an indispensable problem of speculative philosophy to show that its illusion respecting the contradiction rests on this, that we think of man in a different sense and relation when we call him free and when we regard him as subject to the laws of nature as being part and parcel of nature.”¹

Immanuel Kant

“...we are dualists who have two ways of looking at the world: in terms of bodies and in terms of souls. A direct consequence of this dualism is the idea that bodies and souls are separate. And from this follow certain notions that we hold dear, including the concepts of self, identity, and life after death.”²

Paul Bloom

The problem that inspired this thesis is a problem that each person who reflects upon all we know about the world today must at some point encounter. It is the problem of the place of human beings in the world described by science. As someone who has long been interested in the problem of consciousness,³ I am intrigued by the thought of irreducible aspects of human beings that seem to transcend all that we are capable of explaining on the level of objectifying scientific language. The explanatory gap that stands before us seems to be an unbridgeable abyss, until we step out from our usual thinking and reflect upon the way we grasp the world and us within it.

¹ Kant, I. (2004). *Fundamental principles of the metaphysic of morals* [EBook #5682]. (T. K. Abbott, Trans.). Retrieved from <https://www.gutenberg.org/files/5682/5682-h/5682-h.htm>, p. https://www.gutenberg.org/files/5682/5682-h/5682-h.htm#link2H_4_0012

² Bloom, P. (2004). *Descartes' baby: How the science of child development explains what makes us human* [Adobe Digital Editions version]. ISBN 9781446473627, p. 162.

³ Košov, M. (2014). *Modern theories of consciousness and the elusiveness of subjectivity* (Master's thesis). Available from <https://is.cuni.cz/webapps/zzp/detail/136799/>

I decided to refer to the feeling of certain *explanatory gap* that stems from the specific way how people view themselves as persons in the world as *folk dualism* in the thesis, even though this term might have many different connotations. I wish to point out at the very beginning that the reason why I call this conceptual complex “dualism” will become clear when I describe my main methodological tools and findings they provide in the following chapters.

The method by which I approach this problem was not consciously determined in advance. It has organically arisen from my gradual uncovering of different aspects of the problem and discovering inner connections between seemingly unrelated topics I was focusing on in the last several years of my studies (in my papers and during my research in experimental philosophy). That is why my aim is not to provide a review of the historical development of this topic or an exhaustive list of all possible solutions by other authors. Instead, I wish to show the reader the indices as they gradually appeared in front of me and started to form a complex picture, all roofed by the main ideas that originally inspired me.

Throughout the thesis I address numerous different approaches towards the problem or aspects of the problem by various authors in an attempt to show my conclusions on the background of lively current discussion. The main reason for this approach is the fact that in my thesis I combine theoretical and empirical dimensions, with constant development and ongoing discussion being especially characteristic of the latter. The incorporation of the empirical approach into the thematization of the problem of a place of *persons* in the world described by science should be one of the main novelties this thesis is hoping to bring.

In the following passages of the introduction I would like to reveal my main tools: my theoretical starting point and the basis of my empirical methodology. Firstly, I will touch the basic ideas that inspired me to explore the topic and the authors who formulated the problem and its solution in the way that is especially pregnant for what I wish to demonstrate in the thesis. Secondly, I will introduce the empirical side of my method – experimental philosophy, and try to make clear how I understand the value of its contribution to philosophy and to the topic presented in this thesis. After introducing these two pillars I will be able to provide further sketch of the composition of my thesis.

Two realms, two images

My fascination with the problem of the gap between the folk⁴ and scientific concept of a person was strengthened even further after I encountered the thoughts of Donald Davidson and Wilfrid Sellars.⁵ Their conceptions are the basic inspiration for and the thread that stretches throughout my own attempt to get at the problematics of what is the essential aspect of each person - human being, and why does this aspect seem to escape the grasp of scientific approach.

In his essay *Mental Events*, Davidson explains his view called *anomalous monism*. He starts with reference to Kant and the problem of “*reconciling freedom with causal determinism*”. This is his inspiration for how the problem should be handled. For Davidson, mental events are both anomalous and dependent on causal order, just as “*freedom and natural necessity in the same human actions*” are both undeniable for Kant.⁶ Davidson’s aim is to show that if we look at the problem in its full complexity, we will see the apparent contradiction between these facts disappear. He agrees with the assertion about the identity of the mental and physical events⁷ (thence *monism*), but he refuses the view “*that mental phenomena can be given purely physical explanations*” (thence *anomalous*). He admits supervenience of the mental upon the physical, but he refuses its “*reducibility through law or definition*”. Davidson believes that it is not possible “*to reduce moral properties to descriptive*,” just as we cannot reduce truth to syntax.⁸ The key to reconciling the apparent paradox lies in the fact that laws are based on particular kind of linguistic descriptions of the events in question, and the way in which mental events are

⁴ Meaning natural, everyday way we view persons, without deeper philosophical and scientific reflexion. This term will become clearer in the passages where I describe the basic features of experimental philosophy.

⁵ Both thinkers were inspired by Immanuel Kant.

⁶ Davidson, D. (2001). *Essays on actions and events* (2nd ed.). New York: Oxford University Press, p. 207. (Quote from Immanuel Kant: *Fundamental principles of the metaphysics of morals*, 75-6. In the online version cited above: “*between freedom and physical necessity of the same human actions*”, p. https://www.gutenberg.org/files/5682/5682-h/5682-h.htm#link2H_4_0012)

⁷ Ibid., p. 209.

⁸ Ibid., p. 214.

described differs radically from the way in which we describe physical events.⁹ The clarification of the anomalous character of the mental rests upon the “*holism of the mental realm*” – “*Beliefs and desires issue in behaviour only as modified and mediated by further beliefs and desires, attitudes and attendings, without limit.*”¹⁰ In other words: “*Just as we cannot intelligibly assign a length to any object unless a comprehensive theory holds of objects of that sort, we cannot intelligibly attribute any propositional attitude to an agent except within the framework of a viable theory of his beliefs, desires, intentions, and decisions.*”¹¹ Both mental and physical events have their own respective realms within which we can talk about them meaningfully.

The key reason why we are able to effectively use any framework is the fact that we look for coherence when we characterize an item within the framework. In the case of mental realm, we expect persons to be consistent – we expect their propositional attitudes to come from a proper “*place in the pattern*” of other propositional attitudes. A *person* is someone who exhibits this consistence.¹² We need a background theory that will enable us to ascribe meaningfulness to the person’s words and propositional attitudes: “*In our need to make him make sense, we will try for a theory that finds him consistent, a believer of truths, and a lover of the good (all by our own lights, it goes without saying).*”¹³ Davidson believes that the main role in this theory is fulfilled by the “*constitutive ideal of rationality*”. What makes mental realm anomalous with regard to the physical realm is the fact that we think of human beings as *rational animals*.¹⁴ In the end, this anomalism of the mental is a “*necessary condition*” for regarding a human agent as autonomous being. We don’t appeal to causality based on physical law when describing his free actions. Instead, we “*appeal to his desires, habits, knowledge and perceptions.*”¹⁵

It is important for specifying the role Davidson’s thoughts play in my thesis to make clear what is the cornerstone of his anomalous character of mental states. I

⁹ Ibid., p. 215.

¹⁰ Ibid., p. 217.

¹¹ Ibid., p. 221.

¹² Ibid., pp. 221f.

¹³ Ibid., p. 222.

¹⁴ Ibid., p. 223.

¹⁵ Ibid., p. 225.

have already suggested in the previous short review that it is *the ideal of rationality*. As Jaroslav Peregrin suggests in his interpretation of Davidson, when we perceive human beings “*through the prism of rationality*”, we aim at an *ought*, and not simply at an *is* typical for descriptive methods applicable to animals or artefacts. This normative dimension is focused on rationality, meaning that each person should behave in an *ideally rational* way. This is a basis for an effective reciprocal interaction between agents who interpret each other.¹⁶ That being said, there still might be a need to search for further views that will go beyond the formulation of the ideal of rationality in order to capture the essential aspect of the folk concept of a person.¹⁷ This might become more clear when we turn to the conception of Wilfrid Sellars.

In his essay *Philosophy and the Scientific Image of Man* Sellars addresses the problem of the concept of a person in a way that resembles Davidson’s anomalous monism mainly in the conclusion that the framework of persons is irreducible to purely scientific description.¹⁸ Instead of the *mental* and the *physical* he speaks about *two images* - “*the manifest and the scientific images of man-in-the-world*”, that stand against each other face to face with the challenge of *fusing them into one vision*.¹⁹ The manifest image isn’t simply some pre-reflective framework, quite the contrary - it is the framework that is a result of a reflective act that made man actually become man, “*the framework in terms of which man came to be aware of himself as man-in-*

¹⁶ Peregrin, J. (2018). Davidson and Sellars on “two images”. *Philosophia*, 46(1), 183-192, p. 191.

¹⁷ Even though rationality is central for Davidson, I believe that moral dimension which is important for me in this thesis is implicitly present in this conception. Mental realm is a realm of moral interactions, and we cannot “*reduce moral properties to descriptive*” (*Essays on actions and events*, p. 214). What is more, in his description of the background theory he writes that “*...we will try for a theory that finds him consistent, a believer of truths, and a lover of the good...*” (p. 222). Despite his focus on rationality, it seems that an implicit goal of a rational free action of a human agent is the good. The content of this good remains unspecified, though.

¹⁸ Sellars, W. (1963). *Science, perception and reality*. London: Routledge & Kegan Paul Ltd., p. 39.

¹⁹ *Ibid.*, p. 4f.

the-world.”²⁰ The manifest image itself bears certain characteristics of a scientific image – it encompasses certain discipline, critical thinking and uses “*correlational induction*” as its method. The crucial difference between the manifest and the scientific images lies in the fact that the scientific image operates with “*the postulation of imperceptible entities, and principles pertaining to them, to explain the behaviour of perceptible things*.”²¹

The manifest image can be viewed as “*the refinement of the 'original image'*” that could be characterised by universal ascription of personhood to all things. With the ascend of the manifest image, the status of being a person was taken away from everything except for *human* persons (certain kind of “*de-personalization*”).²² As I understand Sellars, by determining the category of persons in such a narrow way, humans became aware of themselves as persons in contrast to the rest of the animals and the rest of the world – they started to exist *as persons*. Analogical to Davidson, the question for Sellars remains: has a human being *as a person* a chance of surviving while being engulfed by the scientific image, “*conceived in terms of the postulated objects of scientific theory?*”²³

As Sellars notes, there are many scientific images of man. One image is characteristic of physics, another of biochemistry, another of social science, etc. All these images contrast with the “*sophisticated common sense*”, with the way humans see themselves “*at the properly human level*”- the manifest image.²⁴ Sellars arrives at certain obstacles that stand in a way of unifying these two images. One of the obstacles hides in the “*intrinsic character*” of sensations. In the attempt to “*fit together the manifest sensation with its neurophysiological counterpart*” we encounter the incompatibility of “the ultimate homogeneity of the manifest image” and “the ultimate non-homogeneity of the system of scientific objects.”²⁵ With the present state of science we are not able to reconcile these two realms. We would have to “*penetrate to the non-particulate foundation of the particulate image*” in

²⁰ Ibid., p. 6.

²¹ Ibid., p. 7.

²² Ibid., p. 10.

²³ Ibid., p. 18.

²⁴ Ibid., p. 20.

²⁵ Ibid., p. 36.

order not to lose the ultimately homogenous realm of sensations from the picture.²⁶ This is akin to the hard problem of consciousness, which is, no doubt, necessary to address in order to map the whole problem of the place of man as a person in the world. I will do so in the last chapter of the thesis.

The obstacle that is much more challenging for Sellars (as well as for me in the context of this thesis), is connected to the very fact that human beings are *persons*, beings with ethical, logical, and other standards, “*responsible agents*” who are at the same time drawn by many passions, wishes, and impulses that place them in front of often very difficult *free* choices.²⁷ Normativity connected to personhood is the ultimately irreducible aspect of a human being. The “*irreducibility of the personal is the irreducibility of the 'ought' to the 'is'.*” Most importantly, a human being as a person is a member of a community of beings capable of “*meaningful discourse*”. The community sets the standards of “*'correct' or 'incorrect', 'right' or 'wrong', 'done' or 'not done'*”, the “*common intentions*”.²⁸ The only solution with regard to the attempt to reconcile the images is to *join* “*the conceptual framework of persons*” to the scientific image and to enrich the scientific framework with “*the language of community and individual intentions*”.²⁹

As Peregrin sums up, while in Davidson the *ought* is oriented towards the ideal of rationality, in Sellars it pertains “*to certain kind of solidarity*” with the community. The community draws us to a common goal, towards respecting the rights and duties defined by it. With each of our autonomous actions we head either towards praise or condemnation by the society we belong to.³⁰ In Peregrin’s words: “*While Davidson’s concept is closer to an instrumental concept, particularly as it is interwoven with belief-desire psychology, Sellars’s concept appears to be closer to the broad, Kantian concept, according to which rationality and rules (which constitute the “rights” and “duties”) are two sides of the same coin.*”³¹

²⁶ Ibid., p. 37.

²⁷ Ibid., p. 38

²⁸ Ibid., p. 39.

²⁹ Ibid., p. 40.

³⁰ Peregrin, J., *Davidson and Sellars on “two images”*, p. 191.

³¹ Ibid., p. 190.

Even though Sellars's concept seems to capture the problem more complexly, both thinkers inspired my quest towards the irreducible core of human being defined as a person and will enable me to demonstrate how fittingly it corresponds to what seems to draw the intuitions of the folk – people who depend on common sense and form the community Sellars talks about. In order to do this, I will need my second tool – experimental philosophy.

Experimental philosophy³²

Experimental philosophy (a.k.a. x-phi) is a relatively new methodological approach in philosophy that continues to gain popularity around the philosophical world due to its approach towards philosophical intuitions: x-phi turns to empirical methods in order to gain insight into folk (laymen) intuitions connected to different philosophical concepts and psychological factors that lie behind these intuitions. The methods that x-phi uses, inspired by psychology and social sciences in general, include questionnaires, interviews, some neuroscientific methods, and, most importantly, statistics.

The first studies that can be characterised as experimental philosophy started to appear in American philosophical environment at the beginning of this millennium. The experimental philosophy manifesto was introduced by Joshua Knobe and Shaun Nichols who see experimental philosophy as a natural descendant of a much older tradition. The key aim of x-phi is, in their view, to learn how human mind works. This complex phenomenon which was shaped due to intricate changes throughout human history and which differs culture from culture has always been an important subject for philosophical inquiry.³³

³² This introduction to experimental philosophy is based on the literature review I used in my paper (section "Introduction" – pp. 581-584, and section "Critics" – pp. 592-596) in which I address my attitude towards this methodological approach in greater depth: Jirout Košová, M. (2020). Skúmanie významu experimentálnej filozofie skrze koncept osobnej identity [Exploring the significance of experimental philosophy through the concept of personal identity]. *Filosofický časopis*, 68(4), 581-603. (Part of the grant project GA UK 925416)

³³ Knobe, J., & Nichols, S. (2008). An experimental philosophy manifesto. In J. Knobe & S. Nichols (Eds.), *Experimental philosophy* (pp. 3-14). New York: Oxford University Press, p. 3.

The symbol of experimental philosophy movement is a burning armchair - a symbolic challenge to the traditional (“armchair”) approach based on classical analytical methods and relying on the philosopher’s own rational thinking and intuitions. Experimental philosophers challenge the traditional old-school analytical philosophers to “stand up from the chair” and get in contact with the actual reality, with the outer world, via embracing empirical data that are around us for the taking if we approach the world in the right way and ask the right questions.

The ways of approaching and using this new philosophical tool soon began to part, however. For the sake of simplicity, we can say that there are two main branches of x-phi: the positive and the negative approach. Antti Kaupinnen defines this division as follows:

“(EXPERIMENTALISM -) Armchair reflexion and informal dialogue are not reliable sources of evidence for (philosophically relevant) claims about folk concepts. (EXPERIMENTALISM +) Survey studies are a reliable source of evidence for (philosophically relevant) claims about folk concepts.”³⁴

According to Alexander, Mallon, and Weinberg, positive approaches in experimental philosophy assert that intuitions about philosophical problems are a relevant empirical source contributing to philosophical knowledge. What is important, though, is not to rely solely on the intuitions of professional philosophers, but to work with philosophical intuitions of the folk - people from wider public, who are laymen in philosophy. Another step is to decide what is meant by “philosophical knowledge” - experimental philosophers can either aim at the knowledge about psychological entities in people’s heads, or on non-psychological entities - some kind of philosophical truths in the stronger sense.³⁵

Negative approaches in experimental philosophy are critical towards the aspirations of the positive branch. They believe that intuitions are not a reliable source of philosophical knowledge, and that they are often very dissonant across

³⁴ Kaupinnen, A. (2014). The rise and fall of experimental philosophy. In J. Knobe & S. Nichols (Eds.), *Experimental philosophy (Volume 2)* (pp. 3-29). New York: Oxford University Press, p. 5.

³⁵ Alexander, J., Mallon, R., & Weinberg, J. M. (2014). Accentuate the negative. In J. Knobe & S. Nichols (Eds.), *Experimental philosophy (Volume 2)* (pp. 31-50). New York: Oxford University Press, p. 35.

different groups of people. The negative x-phi approach focuses on uncovering numerous factors that influence the resulting intuitions of people. How can positive approach defend one concrete intuition in the face of empirical results pointing to the fact that people are subjects to innumerable conflicting intuitions on the same subject?³⁶

The criticism of the positive approach in experimental philosophy is reasonable. Antti Kauppinen asserts that the questionnaire method commonly used in x-phi has little chance of uncovering “robust intuitions”, giving us a complete picture of factors that drove the respondents to answer our questions the way they did. Robust intuitions could only come from a competent user of the concept in question, who has enough knowledge and reflexion to give competent answers. What we get from the “folk” (respondents from wider public who are laymen in philosophy) instead are “surface intuitions”.³⁷ What is more, questionnaire studies using limited scale of possible answers (yes or no, or various kinds of Likert scale) have little chance to provide us with answers that reflect the real opinions (the respondent is partially forced to answer in a certain way) and reveal the motivations that lie behind the respondent’s answer. Their interpretation is therefore often very shaky. Method of dialogue, on the other hand, doesn’t fulfil the objectivity that is characteristic of the scientific rigour that x-phi aspires to.³⁸

Kauppinen doesn't accept the negative branch of x-phi either. Negative approach of experimental philosophy criticises philosophers relying on their own reflexion and philosophical dialogue. Kauppinen believes that philosophical dialogue is the right approach, since philosophers are constantly being trained in proper grasp of philosophical concepts.³⁹

Ichikawa provides an overview of the way x-phi approaches philosophical intuitions. Despite some critical points he still sees negative approach in experimental philosophy as an important tool helping us in reflexion upon philosophical thinking. X-phi methods are useful in revealing different biases hidden

³⁶ Ibid., pp. 34f.

³⁷ Kauppinen, A., *The rise and fall of experimental philosophy*, pp. 9f.

³⁸ Ibid., pp. 16f.

³⁹ Ibid., pp. 20-24.

within philosophical thinking and should be seen as an integrative part of philosophical work, not as a movement that goes against philosophical tradition.⁴⁰

Alexander, Mallon, and Weinberg also believe that experimental philosophy in its negative form has a great potential of revealing important factors behind folk intuitions (influence of culture, concreteness and abstractness of the case, emotional valence, etc.).⁴¹ At the same time, they criticise current failures of its insufficient methods and suggest that experimental philosophy can survive as a useful approach within philosophy only if it becomes more scientifically strict, e.g. via approximating experimental psychology in the level of its scientific qualities. Integrating philosophy into the scientific discipline thus described is, however, a great challenge.⁴²

I agree with the critiques of the positive branch of x-phi and see the negative alternative as the meaningful choice. At the same time, I am aware of the weak spots that are characteristic of the x-phi experimental studies: the restricted answer options and vignettes are an obstacle to uncovering the folk concepts in their full depth and provide only a fuzzy idea of folk intuitions. However, I believe that the idea that from one study we can infer that people in general have such and such concept or intuition is deeply erroneous. Any generalisation based on one or just a few studies goes against the basic scientific principles. As experimental philosophy develops, its authors realise this more and more and the experimental philosophy practice is currently far from naive generalisation from one study. It is rather a dynamic dialogue between proponents of different interpretations, numerous replications and no definite conclusions.⁴³

It is important to make clear that the very subject of focus for experimental philosophy - folk concepts and intuitions - is very fuzzy and hard to grasp.⁴⁴ I fully

⁴⁰ Ichikawa, J. J. (2016). Intuitive evidence and experimental philosophy. In J. Nado (Ed.), *Advances in experimental philosophy and philosophical methodology* (pp. 155-173). Bloomsbury. Here I refer to the online version retrieved from <https://philarchive.org/archive/ICHIEAv1>, pp. 11f.

⁴¹ Alexander, J., Mallon, R., & Weinberg, J. M., *Accentuate the negative*, p. 34.

⁴² Ibid., pp. 44ff.

⁴³ Jirout Košová, M., *Skúmanie významu experimentálnej filozofie skrze koncept osobnej identity*, p. 598.

⁴⁴ Ibid., p. 599.

agree with the view that folk intuitions are “theory-lite”.⁴⁵ This means that the folk have intuitions that are not necessarily logically impenetrable and built on a complex metaphysical theory. On the contrary, folk concepts are vague and blurry, because folk don't need them to be different. As Wittgenstein notes in his *Philosophical investigations* on a blurry photograph: “Isn't the indistinct one often exactly what we need?”⁴⁶ It is thus nonsensical to expect that experimental philosophy could reveal sharp contours of any folk concept. Instead, its methods adapt to its subject: by numerous studies, always slightly different vignettes, different answer possibilities, experimental philosophers uncover folk intuitions step by step in the midst of continuous debates, mutual friendly critique, and inspirations for new and new studies.⁴⁷

Experimental philosophy is a process via which we come to understand how philosophical intuitions were born in the minds of real persons, people untouched by philosophical training, people who are still in touch with the real human world. What would remain of philosophy if it didn't see the loss of touch with the real human life as its gross failure?⁴⁸ This especially applies to the field of moral philosophy.

I believe that the topic of my thesis cannot be properly covered without turning attention towards folk intuitions. When we ask about how persons understand themselves as part of the world, it would be a mistake to leave the voice of the folk (the *actual persons*) out of the question. The whole problem turns around how *people* normally conceptualize persons, not around how professional philosophers arrive at the right definitions and concepts. In what follows I wish to show that when we

⁴⁵ Nahmias, E., & Thompson, M. (2014). A naturalistic vision of free will. In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy* (pp. 86-103). Abingdon: Routledge, p. 89.

⁴⁶ Wittgenstein, L. (1958). *Philosophical investigations* (2nd ed.). Oxford: Basil Blackwell, p. 34 (§71).

⁴⁷ See e.g. Machery, E., & O'Neill, E. (Eds.). (2014). *Current controversies in experimental philosophy*. Abingdon: Routledge.

⁴⁸ Jirout Košová, M., *Skúmanie významu experimentálnej filozofie skrze koncept osobnej identity*, p. 600.

complement theoretical argumentation with empirical tools that x-phi provides,⁴⁹ we will arrive at a much richer and complex picture of the problem at hand.

Bearing in mind my main tools, I will now turn to covering different aspects of the problem in the four main chapters. In the first chapter I will look at the concept of *self* – the more general concept that introduces the whole topic from a wider perspective. By introducing and confronting different notions of the concept of self I will point to the way in which I wish to approach the whole problem of the “dualism” between the two realms and two images.

In the second chapter I will address the first of the three main aspects or sub-concepts of the concept of self⁵⁰ that are crucial for purposes of the thesis, namely the *causal agent* and the related problem of *free will*. I will demonstrate how the theoretical background of my solution of the problem (parallel to the first chapter) stemming from the sources of my inspiration fits and is complemented by the empirical results that are available thanks to experimental philosophy.

The third chapter is the central chapter of the thesis because it turns to the second and in my view the most important aspect of the self, namely the *essential self*. In this chapter I will turn to the *normative* dimension of the concept of self and show that it is crucial for understanding the special status of persons in the world. The chapter is based on the results of empirical studies that address the folk concept of personal identity via the notion of *true self* and its connection to the folk concept of *soul*. By introducing different existing studies (including my own studies) I will demonstrate that interpersonal relationships and morality connected to life in human community are the very cornerstone of the folk concept of self and the way people

⁴⁹ When referring to the results of x-phi studies, I decided not to include statistical details and exact results in order to make the text of the thesis more reader-friendly. All the results are available in the publications I refer to.

⁵⁰ The three sub-concepts (*causal agent*, *subject*, and *essential self*) that I found to be the main pillars of the concept of self that I wish to introduce in the thesis are in accord with the sub-concepts addressed by Berniūnas, even though the context in which Berniūnas uses them differs from the context in which I develop the concept of self. See Berniūnas, R. (2012). Folk concept of ‘a person’: Structure and warrant. *Problemos*, (Supplementary), 63-77, p. 69.

conceptualize persons as irreducible in the face of scientific explanations of the world.

In the fourth chapter I will address the last of the three main aspects of the self, namely the self as a *conscious subject* of experiences. By thematising and criticising the philosophical problem of consciousness in the face of empirical data I will try to determine its role within folk dualism and show that this role is often overestimated by philosophers. I will show how experimental philosophy problematizes and helps to clarify the role of consciousness in the folk view of persons and points once again to the crucial importance of moral and interpersonal dimension.

In the conclusion, I will further clarify the relationship between the three aspects of the concept of self and show how they work together and form the image of man in the world that escapes scientific approaches and that should never be reduced to or considered less important than the scientific image. Thanks to thus revealed cornerstone of folk dualism we gain a tool that might help us handle moral challenges that stand in front of us in the age of scientific progress and the ascend of future technologies.

1 The Concept of Self⁵¹

This chapter deals with a wider philosophical problem of the *self*. The concept of *self* is tightly connected to human beings viewed as *persons* and beings that have a special place in the world. Self is not only the bearer of our experiences, agency and free will, but also our inner nature and identity. We deal with selves on an everyday basis, yet we face unusual quandaries when asked to identify their real nature. Selves seem to be more than just our bodies, even more than just our minds. They somehow transcend all our properties and provide each of us with uniqueness, unity and certain sense of essence. Where does this specific irreducibility of the *self* originate? By confronting different views of the self, I will try to formulate my own notion of this concept and thereby introduce this crucial term within the basic context that will interest me in the thesis.

Renatas Berniūnas notes that concepts of *causal agent* (concerning free will), *subject* (concerning consciousness) and *essential self* (concerning personal identity) are the most important components of the so-called “*minimal self*”. People view certain bodily features and psychological states as natural parts of the self, yet they also have this “*minimal conception of the self*” that is transcendent with regard to the person’s body and psychology.⁵² I will address this narrow concept and its sub-concepts in detail in the subsequent chapters. Like Berniūnas, I have arrived to the conclusion that these three concepts are essential building blocks of the conception man has “*of himself as man-in-the-world*,”⁵³ even though I will specify the content of these concepts for my purposes. Before turning to the concrete sub-concepts one by one, I will introduce the problem of the self from a wider point of view in this chapter.

⁵¹ This chapter is mostly based on my paper (published in Slovak language): Košová, M. (2014). Skutočná podstata *ja* [The true nature of the self]. *Pro-Fil*, (Supplementary), 50–64. Available from <http://www.phil.muni.cz/journals/index.php/profil/article/view/998>

The study is set within the new context of the thesis; thus, some revisions were necessary. However, it would not be reader-friendly to cite the paper consistently throughout the chapter. To access the original paper, please visit the link cited above.

⁵² Berniūnas, R., *Folk concept of ‘a person’: Structure and warrant*, p. 69.

⁵³ Sellars, W., *Science, perception and reality*, p. 6.

It is not my aim here to provide an exhaustive overview of all possible philosophical views concerning the concept of self. Instead, my method is based on choosing concrete views that will help me illustrate the problem by putting them into confrontation. This method helps me set the direction I wish to follow and stay focused on the aspects of the whole problem that I wish to point to. It is precisely by reflecting upon these views that I became able to see connections between different concepts and thoughts that I present and develop throughout my thesis.

In the first part of the chapter I will introduce theory of Colin McGinn who examines various approaches to the self and finally turns to a position which he calls *transcendental naturalism*. The second part focuses on more scientific analyses of the self from the perspective of Daniel Wegner and Daniel Dennett who view the self as an illusion that serves as a useful mental construction. These three authors should provide an example of possible attitudes towards the problem that especially interest me due to their mutual tension at the one hand, and a basis for a more complex view on the other. In the third part I will return to Donald Davidson's idea of *anomalous monism* which should lead me on a way to formulate such conception of the self that would satisfy some of McGinn's intuitions, include Wegner's and Dennett's conclusions and at the same time avoid the need to invoke transcendental naturalism. In the fourth part I will try to explain my idea of this conception with relation to all the mentioned theories (briefly bringing also Sellars and his pertinent ideas into the picture) and point to its strengths. This should serve us as an illustrative introduction to the following chapters where I will develop the picture of the concept of self and its relation to folk dualism in further details.

1.1 Transcendental naturalism

According to McGinn, the main question about the self is concerning its unity. Self brings unity to conscious states⁵⁴ and by this creates a *centre* which stands on its own, distinguished from all the other conscious centres. To explain this seems to be

⁵⁴ One of the most important aspects of the self is its connection to consciousness. I will turn to the folk concept of irreducible conscious properties and its relation to the problem of folk dualism in the fourth chapter.

extremely difficult – thus we encounter a philosophical problem.⁵⁵ We cannot conceive of selves as some spatial entities, because they are closely connected to consciousness. Self is not only the body. It seems to be more than just this physical object.⁵⁶ This means that we can't perceive selves directly through our senses: "*I can see that that your body is distinct from his, but I cannot in this way see that you are distinct from him, since selves are not perceptually presented in the way bodies are (though they may be perceptible in some derivative way).*"⁵⁷ What is more, part of the philosophical problem with the self has to do with "*the systematic transcendence of the self in acts of self-awareness*" – "I" can never become an object of my focusing on myself, because I will always be there as the *subject* and there is no way how to objectively capture this fact.⁵⁸

In other words, it is obvious that we must approach the self in a special way. It is not some directly observable physical object, and our usual concepts seem to fail when applied to it. The self seems to require some new level of conceptualisation.

The need for a different approach is illustrated by the failure of our tendencies to "domesticate" the self. Some theories might identify the self with the body or with the brain. According to these views there are no persons over and above the bodies, because the self is simply a physical object, just like other objects we know. There is no essence that would transcend the material aspects of the self.⁵⁹

Another way to domesticate the self is to define it as a series of mental states related to each other via "*memory, causal continuity, psychological similarity and so forth.*" Again, the self is nothing over and above the mentioned relations. There is no mysterious entity that transcends particular mental states and the way they relate to each other.⁶⁰

If we accepted the domesticating theories, we would be able to describe and define the self in familiar terms. There would be no need for new and radically

⁵⁵ McGinn, C. (1993). *Problems in philosophy: The limits of inquiry*. Oxford: Blackwell, p. 46.

⁵⁶ Ibid., p. 47.

⁵⁷ Ibid., p. 48.

⁵⁸ Ibid.

⁵⁹ Ibid., p. 50.

⁶⁰ Ibid.

different concepts. However, McGinn believes that these attempts to reduce the self to familiar aspects of reality are actually unsuccessful. He believes that these theories encounter similar problems as those that are characteristic for the mind-body problem.⁶¹

As McGinn points out further, to assert irreducibility of the self doesn't solve the problems either. We need to explain how the self is linked to the "*other things*". It is not obvious at all that self should be independent from everything else in the world.⁶²

Our inability to grasp the self and its place in the world might lead us (and indeed often leads us) towards unwarranted supernaturalism. Viewing the self as a magical entity is only a symptom of our inability to define it in familiar scientific terms. It is a desperate attempt to find some place for the self in the world.⁶³

Another possibility to deal with the problem of the self is eliminativism, but also this approach is doomed according to McGinn. It seems absurd to claim that whenever we speak of persons, we in fact lack any real reference. "*Surely the ontology of persons is rooted deep in our thought and speech; to abandon it would be to abandon something pervasive and useful, to say the least.*"⁶⁴ We simply have to find a way how to do justice to the concept of self as something which points to an essential aspect of our world.

McGinn has his own answer to our problem with understanding the self. It is a theory he calls *transcendental naturalism*. It could be characterised in the following way: "*Reality itself is everywhere flatly natural, but because of our cognitive limits we are unable to make good on this general ontological principle. Our epistemic architecture obstructs knowledge of the real nature of the objective world. I shall call this thesis transcendental naturalism, TN for short.*"⁶⁵ With respect to the self, TN provides us with a possibility to assert the reality of the self and at the same time avoid all the unsuccessful "*positions on the DIME*⁶⁶ *shape*" described above.⁶⁷

⁶¹ Ibid.

⁶² Ibid., p. 52.

⁶³ Ibid., p. 53.

⁶⁴ Ibid.

⁶⁵ Ibid., pp. 2f.

⁶⁶ The positions which view the self as Domesticated, Irreducible, Magical or Eliminated.

McGinn believes that certain properties of persons exist which make persons as such possible and that these properties “*transcend our conceptual resources*”. This means that although domestication, irreducibility, and supernaturalism don’t work as solutions to our problems with grasping the self, there is no reason to accept eliminativism. The fact that our cognitive capacities fail to deal with some questions doesn’t mean that the facts behind the questions don’t really exist. We simply “*cannot assume that the true objective nature of the self is adapted to our given modes of cognition.*”⁶⁸

The main problem, as stated by McGinn, concerns the relation between the self and “its” body (spatially extended physical object) and mind (collection of mental states).⁶⁹ We are prone to view things we encounter in a certain specific way (CALM – combinatorial atomism with lawlike mappings). McGinn describes this mode of thought as suited to subject-matters “*in which an array of primitive elements is subject to specified principles of combination which generate determinate relations between complexes of those elements.*”⁷⁰ This doesn’t apply to the self. The self is not its body, and it is not composed of parts in such a way as the body. It is not simply composed of its mental states either. According to McGinn, the self is something over and above its bodily/brain parts and mental states. That’s why it is misleading to understand the claim that the self *has* a body and a mind “too literally”, along the CALM lines. We simply cannot see the self as an object that consists of physical and mental constituents, because the reality is “*more mysterious and unique than that – hence the existence of a philosophical problem.*”⁷¹

McGinn realises that selves depend on other aspects of the world. Persons emerge from certain specific biological conditions at some point in evolution, but we need to presuppose something “extra” that will enable this peculiar ontological jump. Persons have to be explicable in terms of some natural science, even though we can be cognitively closed in respect to this science.⁷²

⁶⁷ Ibid., p. 53.

⁶⁸ Ibid., p. 54.

⁶⁹ Ibid., p. 55.

⁷⁰ Ibid., p. 18.

⁷¹ Ibid., pp. 55f.

⁷² Ibid., p. 56.

What is more, it seems obvious that selves supervene on the physical – the way molecules in my body/brain are arranged somehow determines the existence of my personhood. If some other body and brain were arranged in the same way, they would yield a person too. This is similar to the claim that mental states supervene on brain processes. McGinn believes that there has to be an explanation for this undeniable dependence.⁷³

McGinn suggests that the self is a natural part of the world, a biological entity which falls under natural principles. However, in the spirit of defence of transcendental naturalism, he adds that the concepts we are able to come up with are insufficient for the task of explaining how the self relates to the rest of the world. We cannot account for the self as a natural thing because our scope of the natural world and its laws and principles is limited. The nature of the self escapes *our* science in principle, and the reason for this is our epistemic boundedness.⁷⁴

The above-mentioned conclusions imply that there exists some “*hidden structure of the self*”. Unity of the self and the way it relates to its body is based on a certain “*inner architecture*”. The trouble is that our conceptual tools are not suited to grasp this “*hidden nature*”. It nevertheless exists – at least we are not warranted to deny its existence by pointing to our inability to reach a proper understanding of it. Some other beings with a radically different cognitive equipment could in principle be able to grasp the essence of the self. The problem of grasping the self doesn’t dwell in our inability to find the right way the self - as we usually perceive it - relates to the body. Rather, we are not able to see the “*internal constitution*” or “*hidden structure*” of the self which makes everything fall into place. There is an “*extreme epistemic gap that separates our conceptions from the objective nature of what they refer to.*” We are adapted to understand physical objects in space but not such things as selves.⁷⁵

As I understand McGinn, he tries to say that the self is not an entity radically different from the physical entities (e.g. bodies or brains) that we usually have no larger problems describing and accounting for. It is only that our scope is too narrow – we are only adapted to properly grasp some types of things in the world.

⁷³ Ibid., pp. 56f.

⁷⁴ Ibid., p. 57.

⁷⁵ Ibid., pp. 58f.

To sum up the main thoughts of McGinn's theory of the self, he believes that there is some hidden structure or essence of the self which underpins its unity and personhood and that this hidden nature is knowable in principle by a certain kind of science which is unattainable from the human perspective. There is no reason to reduce the self to something we can easily grasp or to claim that the self is magical or doesn't exist after all. None of those approaches provide a satisfactory answer for McGinn. For him, the only way to acknowledge the existence of the self and at the same time account for our inability to grasp its essence is to embrace transcendental naturalism thesis.

1.2 Self as a useful construction

McGinn characterised the self as something that brings unity to conscious states and keeps each conscious centre distinguished from all the other conscious centres. He also mentions the close connection of the self to persons and suggests that "*the ontology of persons*" is very natural for us. However, I believe that it would be useful to elaborate the concept of self further. What does it mean to say that the self lies at the root of unification of conscious mental states? Under what conditions do we ascribe personhood to a being? By trying to answer these questions we might be able to better deal with the idea of the hidden objective structure of the self.

Gallagher identifies two important concepts of the self: the "*minimal self*"⁷⁶ that is unextended in time, and the "*narrative self*" that is continuous across time and constitutes a base for personal identity.⁷⁷ The minimal self is experienced as "*a consciousness of oneself as an immediate subject of experience, unextended in time.*" It is connected to the sense of ownership (the feeling that I am undergoing certain experience, that it is my body which is moving etc.) and action (the feeling that I am causing a certain action or a certain thought). The narrative self, on the other hand, is

⁷⁶ The concept of *minimal self* addressed by Gallagher differs from the concept of *minimal self* of Berniūnas mentioned in the introduction to this chapter. Gallagher is concerned with the conscious aspect of the self, whereas Berniūnas refers to a narrow concept of self that is more complex and touches more domains (including the moral domain, as we will see especially in the third chapter).

⁷⁷ Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Sciences*, 4(1), 14-21, p. 14.

a “*more or less coherent self (or self-image) that is constituted with a past and a future in the various stories that we and others tell about ourselves.*” It is extended in time because it includes memories of the past and intentions concerning the future.⁷⁸

Firstly, I would like to point to some interesting aspects of the concept of minimal self as Gallagher defines it. Wegner addresses this topic by referring to the so-called “*homunculus problem*”. It stands for an idea that there is some “*inner executive*” who performs the person’s actions. It is the first and free-willing cause of our actions and no prior cause influences and determines its decisions. This “*renders it an explanatory entity of the first order. Such an explanatory entity may explain lots of things, but nothing explains it.*” Explanatory entity of the first order prevents scientific approach since it stops any other attempt to explain the given phenomena in more detail.⁷⁹ It is a transcendent entity, because it stands at the very border of causal structure of the scientifically understood world. This constitutes an important part of the problem of free will that will be discussed in the next chapter.

As Wegner suggests, the homunculus problem comes on stage especially when it comes to the question of the subject of controlled and automatic processes. The reason for this has to do with the fact that the feeling of control over our actions plays a crucial role in our understanding of the self. People have a strong tendency to view controlled processes as connected to personhood, unlike automatic processes: “*Controlled processes are often seen as conscious, moral, responsible, subtle, wise, reflective, and wilful, not because they are described as such in so many words, but rather because they are what is left when we subtract the automatic processes.*”⁸⁰ This preference for controlled processes might be connected to our bias to spot minds (agents): “*Early in life, we develop the tendency to understand events that are attributable to minds, and to distinguish them from events that are caused by mechanical processes.*” We are prone to expect “*minds behind events*” because of our experience with our own mind. We have a feeling that we control our own

⁷⁸ Ibid., p. 15.

⁷⁹ Wegner, D. M. (2005). Who is the controller of controlled processes? In R. R. Hassin, J. S. Uleman & J. A. Bargh (Eds.), *The new unconscious: Social cognition and social neuroscience* (pp. 19-36). New York: Oxford University Press, p. 20.

⁸⁰ Ibid., p. 21.

actions. This results in “*the further intuition that there is always an agent behind the processes that control human thought and action.*”⁸¹

This could be a basis of the concept of minimal self. The processes that really constitute *our selves* have to be fully experienced as intended. Wilful and intended action is what makes us persons. However, as Wegner is trying to show, the idea of a free willing controller might be an illusion: “*The personal experience of agency is not a good foundation for a science of mind, however, and we must be careful as scientists to appreciate the basis of this feeling rather than to incorporate the feeling in our theories.*”⁸² In order to explore the true nature of the self we have to consider a possibility that the minimal self is only a construction based on the interpretation of our experiences.

According to Wegner, our experience of control over our actions is based on inferences we make about our thoughts and actions, when they are properly timed. “*In essence, the theory suggests that we experience ourselves as agents who cause our actions when our minds provide us with previews of the actions that turn out to be accurate when we observe the actions that ensue.*”⁸³ Conscious thoughts and intentions are especially important for the inference of an agent. It is important to stress that they do not cause the action; conscious thoughts and intentions provide a “*conscious preview*” of an action, and since they are often consistent with the action, they elicit an experience of conscious will⁸⁴ and fuel the inference of a controller.⁸⁵

Wegner concludes that controlled processes (in contrast to automatic processes) give us enough space to infer causal relationships between our minds and our behaviour. These causal inferences are a basis for our creation of the idea of a “*virtual agent*” who is in charge of our actions. Thus this agent or a controlling mind is our construction, not a real entity, and Wegner doesn’t forget to emphasize this point: “*Although this mind is a deeply important construction, allowing us to*

⁸¹ Ibid., p. 22.

⁸² Ibid., p. 23.

⁸³ Ibid.

⁸⁴ Amongst a number of other authors, Sven Walter provides a sophisticated critic of Wegner’s conclusions concerning the illusoriness of free will. See Walter, S. (2014). Willusionism, epiphenomenalism, and the feeling of conscious will. *Synthese*, 191(10), 2215-2238.

⁸⁵ Wegner, D. M., *Who is the controller of controlled processes?*, p. 28.

*understand, organize, and remember the variety of things we find ourselves doing, it is a construction nonetheless and must be understood as an experience of agency derived from the perception of thoughts and actions – not as a direct perception of an agent.”*⁸⁶

Our sense of agency is indeed very useful in the context of our lives as social beings, even though “the agent” isn’t the real causal source of our actions. Thanks to the experience of conscious will we are provided with an “*authorship emotion*”. This is very convenient since it helps us orientate in our own actions. As Wegner notes, the feeling of conscious will helps us to properly label and spot our actions and distinguish them from other agents’ actions or other events in the world. This process of “anchoring” of our actions via the feeling of conscious will goes beyond any rational mental process. We don’t simply infer that we did the action. Rather, we *feel* it. This immediate connection to our own actions helps us remember, organize and connect our actions into a unified picture of our own self – bearer of our identity.⁸⁷

What is more, we can already see how we can move from the concept of *minimal self* to the broader *narrative self*. Sense of authorship becomes integrated in our memories and future authorship becomes an object of our anticipations. Our life becomes a consistent narration of the wider self.⁸⁸ The self is thus to be understood as something that is constructed and develops over time. We infer an agent from our experiences of authorship and as it becomes incorporated in our memories we anticipate future authorship experiences and form a broader picture of “our selves”: “*We become agents by experiencing what we do, and this experience then informs the processes that determine what we will do next.*”⁸⁹

The idea of homunculus may seem very simple and intuitive (after all, it is useful and effective). However, as Wegner suggests, if we want to grasp the real nature of what hides behind the talk about our intentions or the controller, we have to use terms as “*apparent mental causation*” or “*virtual agency*”. Thanks to this approach we can gain new insights into certain phenomena (e.g. multiple personality

⁸⁶ Ibid., p. 30.

⁸⁷ Ibid., p. 30n.

⁸⁸ Ibid., p. 31.

⁸⁹ Ibid., p. 32.

disorder) which seemed inexplicable when looked at through the lens of the notion of homunculus.⁹⁰

All these ideas seem to lead to a conclusion that the most important aspects of the self arise from how we interpret our experiences. It is a construction based on experiences of our thoughts, actions and feelings, and the inferences we make about their causal relationships. Wegner encourages us to abandon for a moment our natural intuitions about ourselves so that we can gain a much deeper insight into our own nature: “*The way the mind seems to its owner is the owner’s best guess at its method of operation, not a revealed truth.*”⁹¹

Daniel Dennett draws on Wegner’s research and develops his own view of the self that focuses on the emergence of the *narrative self*. He suggests that it is obvious from the way we talk about *our* decisions that we presuppose some kind of a “centre” or “headquarters”. Why are we susceptible to fall for this “*illusion*”? Dennett suggests that the “*illusion of such an ultimate centre*” is a result of “*the idea of the self as a unitary and cohering point of view on the world*”. The idea is strengthened by our “*preoccupations with our responsibility*” and in search of an answer to the question ““*Did I do that?*”” we produce “*something like a geometrical construction in search of interpretation.*”⁹²

This “geometrical point” could be a basis for the concept of *minimal self* which gradually transforms into the *narrative self*. We are looking for an agent who is the one in charge of our actions. This agent is associated with personhood, and we identify with it in much stronger sense than with any other unconscious or automatic process. Dennett, just as Wegner before, is trying to show that it is a construction and a mere result of our interpretation.

It is, nevertheless, a useful construction. It is important to find some principle of unity within ourselves and to be able to view one’s self as distinct from any other self. As Dennett points out, we observe this principle also in biology of primitive organisms. Those organisms behaving in such a way that they managed to preserve their homeostasis got a chance to replicate. Their success depended on how well the

⁹⁰ Ibid.

⁹¹ Ibid., p. 33.

⁹² Dennett, D. C. (1984). *Elbow room: the varieties of free will, worth, wanting*. London: Clarendon Press, p. 78.

particular homeostasis ensured self-preservation. With further self-preservation and self-replication the first “interests” were formed.⁹³ An organism capable of avoiding its own decomposition “*brings with it into the world its own “good”*.”⁹⁴ It isolates itself from the world in a sense that it “sees” everything else in terms of how this *individual* - the particular organised lump of biological matter – is affected by it. A specific individual point of view on the world emerges.

In contrast to simple biological organisms, people have a certain specific trait which characterises their selves: people constantly present and represent themselves, to others and to themselves, using “*language and gesture, internal and external*.” Our environment is very specific in that it contains *words*. We use words in various ways, “*weaving them like spiderwebs into self-protective strings of narrative*.” We protect, control and define ourselves by “*telling stories*”. We are especially occupied by designing the story “*about who we are*”. We tell this story to other people and to ourselves and thus “*our narrative selfhood*” is formed. Dennett puts it quite succinctly when he writes: “*Our tales are spun, but for the most part we don’t spin them; they spin us*.”⁹⁵ This is how our *narrative self* comes into existence.

Unity of the self is a result of the practice of telling the stories about us. Narration encourages the audience to see an “unified agent” behind different words and stories, “*in short, to posit a center of narrative gravity*.” This centre is only an abstraction, not a concrete physical object or a part of the brain, but it is very helpful since it provides a useful simplification. Whatever we encounter, it can be seen as having an owner and a centre from which it arises.⁹⁶ Thanks to the construction of the narrative self – the centre of narrative gravity – we are able to orientate much easier in the world of complex beings.

Alongside Wegner, also Dennett emphasises the fact that the self is not to be seen as something solid and objective. Rather, it is a result of social processes, and it depends on “*the web of beliefs that constitutes it*”.⁹⁷ The self is an abstraction based on a large number of “*attributions and interpretations*”, including those which we

⁹³ Ibid., p. 22.

⁹⁴ Ibid., p. 23.

⁹⁵ Dennett, D. C. (1991). *Consciousness explained*. Boston: Back Bay Books, pp. 417f.

⁹⁶ Ibid., p. 418.

⁹⁷ Ibid., p. 423.

make about ourselves. These attributions and interpretations make up “*the biography of the living body whose Center of Narrative Gravity it is.*” It is obviously very useful for an agent “to have a self” - to be able to distinguish oneself from all the other things.⁹⁸ Centres of gravity are “magnificent *fictions*”⁹⁹, and to understand the centre of gravity which is a self means to accept a naturalistic explanation of how the brain creates “self-representations” and provides the body with a self capable of responsible conduct.¹⁰⁰

In other Dennett’s words, the self is sort of a “*user illusion*”, a useful simplification similar to those we are familiar with as computer users. All those user-friendly “clicks and drags” and sound effects don’t show us how the computer really works – the whole complicated net of mechanisms is hidden behind simplifications which enable the users interact with the computer on an intuitive level, using the users’ natural abilities to perceive the world and act accordingly. In the complex social environment of complex beings each human needs some kind of “subsystem” or program that is created to enable us a smooth interaction with other persons. This subsystem is the self, a concept that provides us and others with a simplification, “*a limited, metaphorical outlook*” on the otherwise complex processes happening in our brains.¹⁰¹

The illusion of self helps each person to track his or her past and future intentions across time. The “centre of narrative gravity” enables me to have “*a means of interfacing with myself at other times.*”¹⁰² This gives the self its unity – whatever I do, I know that the action is mine. I don’t fall apart into a series of unrelated actions. All the actions are pulled towards one centre. This enables us to live in an “orderly world” of responsible persons.

According to the analyses of both *minimal* and *narrative self* introduced above we can conclude that the self could be seen as a mere construction from a certain perspective. The idea of the *minimal self* (an agent or a controller) arises as a result of our interpretation of the action causation. The *narrative self* is constructed

⁹⁸ Ibid., p. 427.

⁹⁹ Ibid., p. 429.

¹⁰⁰ Ibid., p. 430.

¹⁰¹ Dennett, D. C. (2003). *Freedom evolves*. New York: Viking, p. 249.

¹⁰² Ibid., p. 253.

as a useful “user illusion”, a simplification which helps us interact with other people and stay in contact with our own selves at different times. The next step will be to confront Wegner’s and Dennett’s thoughts with McGinn’s conception and to look for a possible new alternative by getting inspiration from yet another thinker – Donald Davidson.

1.3 Anomalous monism

McGinn suggested that we don’t seem to be able to come up with a satisfying theory of the self because we are cognitively limited with regard to its real essence. There is some objective structure of the self which accounts for all its unique qualities, but our science is not capable of revealing it. On the other hand, Wegner and Dennett claim that the self is illusory – it is simply our construction, a mental concept which helps us define ourselves and our boundaries and enables us to interact with each other. If we looked for something objective that “correlates” with the self, we would probably find a bundle of brain activities that play role in interpretation of action causality and in self-representation. There is apparently no hidden objective structure of the self.

It’s not a surprise that it seems very unnatural to think about the self in terms of brain processes. Wegner and Dennett would also agree that the self is not simply reducible to a physical body. However, this doesn’t have to imply that the self is something more than a concept and construction of our minds. McGinn refuses eliminativism, but I believe that seeing the self as a mental concept does not eliminate it. Rather, it moves it to a “different level” – a level of mental entities which has its own undeniable existence. McGinn doesn’t seem to consider this alternative, and that’s why he invokes transcendental naturalism and asserts that the self has indeed some specific objective structure.

In my opinion, there is another alternative apart from eliminativism and transcendental naturalism. On the one hand we don’t have to claim that the self simply doesn’t exist, and on the other hand we don’t have to ascribe to it a hidden objective nature. It is enough to say that the self is a construction of our mind - a mental entity, and that this entity as such cannot be reduced to any physical object or bundling of neural processes, etc., precisely because it belongs to the realm of mental

entities which has its own rules and logic. This view is inspired by Donald Davidson's *anomalous monism*.

Davidson has a specific view of the place of mental events in the world: "*Mental events such as perceivings, rememberings, decisions, and actions resist capture in the nomological net of physical theory.*"¹⁰³ He suggests that even though "*at least some mental events interact causally with physical events*" and "*events related as cause and effect fall under strict deterministic laws*", mental events cannot be "*predicted and explained*" by any "*strict deterministic laws*". It seems obvious that if we hold all the three principles we face a contradiction. Davidson tries to show that the contradiction is only apparent.¹⁰⁴ The following passages are my attempt to show that something similar may apply in the case of the self.

When Davidson is trying to specify what he refers to when he speaks about the mental, he mentions mental verbs. These verbs "*express propositional attitudes like believing, intending, desiring, hoping, knowing, perceiving, noticing, remembering, and so on. Such verbs are characterized by the fact that they sometimes feature in sentences with subjects that refer to persons...*"¹⁰⁵ It seems quite obvious to me that the mentioned verbs are very closely connected to the concept of self. Both minimal and narrative self as described above are in accord with the Davidson's list of mental verbs.

Anomalous monism – the position defended by Davidson – partially resembles materialism. He claims that "*all events are physical*", but at the same time he denies the idea that "*mental phenomena can be given purely physical explanations*" and that "*there are psychophysical laws*". These views are in accord with the idea of supervenience - "*the view that mental characteristics are in some sense dependent, or supervenient, on physical characteristics.*"¹⁰⁶

It is very important to mention that, according to Davidson, allowing for supervenience doesn't imply "*reducibility through law or definition*". Moral properties will probably never be reduced to descriptive, and truth cannot be reduced

¹⁰³ Davidson, D., *Essays on actions and events*, p. 207.

¹⁰⁴ Ibid., pp. 208f.

¹⁰⁵ Ibid., p. 210.

¹⁰⁶ Ibid., p. 214. Note that also McGinn defends the idea of supervenience.

to syntactical properties.¹⁰⁷ Even though mental and physical events are causally connected and all mental events are in fact physical (there is an identity between them), it might be impossible to formulate laws which apply to their connection. As Davidson explains: “*Causality and identity are relations between individual events no matter how described. But laws are linguistic...*” What is important here is the *description*. What makes events mental is the *way we describe* them.¹⁰⁸ As soon as we apply physical description to them, we can formulate a law, but we lose the mental events *as such*. We are simply “*changing the subject*”. We stray away from the “*vocabulary of the propositional attitudes*” which is essential to the mental.¹⁰⁹ Mental events form a realm distinct from the physical realm thanks to the fact that the concepts and vocabulary which we use to describe them differ radically from each other.

No matter how hard we try, we will never be able to describe some mental event using purely physical vocabulary. This explains, for example, our trouble with behaviourism: “*we always find a need for an additional condition (provided he notices, understands, etc.) that is mental in character.*” We can properly account for beliefs and desires only by referring, “*without limit*”, to further propositional attitudes. Davidson speaks here of the “*holism of the mental realm*”. Since every propositional attitude depends on the net of a number of other propositional attitudes, the mental realm displays an “*autonomy*” and “*anomalous character*”.¹¹⁰ This is in perfect accord with the interpretation of experimental results that I introduce in the next chapter on free will: I will show that the folk believe that free will is compatible with certain versions of deterministic scenarios as long as the person’s mental states such as beliefs and intentions play role in the process of forming an action.

Now we are getting to the reason why there are no strict laws connecting the mental and the physical. We can only form “*lawlike*” or “*nomological statements*” by joining predicates of the same type, or, as Davidson explains, “*predicates that we know a priori are made for each other*”. To illustrate the point, Davidson refers to the following example: “*‘Blue’, ‘red’, and ‘green’ are made for emeralds sapphires,*

¹⁰⁷ Ibid.

¹⁰⁸ Ibid., p. 215.

¹⁰⁹ Ibid., p. 216.

¹¹⁰ Ibid., p. 217.

and roses; 'grue', 'bleen' and 'gred' are made for sapphals, emerires, and emeroses." As I understand Davidson, predicates are made for each other when they have the same logic and they are formed under the same rules. Mental and physical predicates don't have the same logic and they fall under different rules. They "*are not made for one another.*" There cannot be strict psychophysical laws because statements about the connection between the mental and the physical resemble statements like "*All emeralds are grue*", whereas a strict law would be stated in the form of "*All emeralds are green*".¹¹¹

We can at best form some generalizations about the relationship between the mental and the physical. A law which lies behind their relationship can be formulated "*only by shifting to a different vocabulary.*" Davidson calls these kinds of generalisations *heteronomic*. For generalization to be "*homonomic*", they would have to lead to such a formulation of a law which "*draws its concepts from a comprehensive closed theory.*"¹¹² The physical and the mental realms, however, fall under different theories which are built on different concepts.

In order to be able to treat someone as a person, we have to move within the right conceptual framework. We have to "*discover a coherent and plausible pattern in the attitudes and actions of others*" and "*the attribution of mental phenomena must be responsible to the background of reasons, beliefs, and intentions of the individual.*" On the other hand, when we talk about the physical reality, we don't refer to propositional attitudes. We operate with concepts like "*physical change*" and "*conditions physically described.*" This once again points to the fact that the mental and physical conceptual realms have "*disparate commitments*". If we want to keep the commitments untouched, we have to face the fact that there simply "*cannot be tight connections between the realms*".¹¹³

Davidson agrees with the view that we can know that mental events are identical with certain physical events (thanks to heteronomic generalizations), but even if we came up with a physical description of the whole world and its history, this wouldn't enable us to "*predict or explain a single mental event (so described, of*

¹¹¹ Ibid., p. 218.

¹¹² Ibid., p. 219.

¹¹³ Ibid., pp. 221f.

course).”¹¹⁴ Mental realm has its own logic and, as was mentioned before, trying to describe it using physical language would amount to changing the subject.

When we operate within the mental realm we are interested in specific explanations. When we speak of actions of a free agent, we explain them by turning to “*his desires, habits, knowledge and perceptions. Such accounts of intentional behaviour operate in a conceptual framework removed from the direct reach of physical law by describing both cause and effect, reason and action, as aspects of a portrait of a human agent. The anomalism of the mental is thus a necessary condition for viewing action as autonomous.*”¹¹⁵

I believe that thanks to anomalous monism we get an explanation of our troubles with finding a proper place for the mental events and for persons in the world. We have to realise that the concepts we use in the context of mental realm differ from the way we think about the physical events. This is obviously very useful in everyday life and makes sense in the context of evolutionary psychology. Anyway, thanks to anomalous monism we can see the mental realm as both natural and irreducible to the physical while we don’t need to invoke transcendental naturalism. In what follows I will attempt to explain this assertion further.

1.4 Self and the two conceptual realms

McGinn certainly managed to show that the self cannot be simply domesticated, reduced or eliminated. Suggesting that it is supernatural or irreducible doesn’t lead very far either: we know that the self relates to its mind and body and that it supervenes on the physical reality. It is apparently a natural part of the world. However, I believe that McGinn doesn’t correctly identify the reason why the self defies simple physical explanation and the CALM approach (the approach we apply to most of the other phenomena in the natural world). He turns to transcendental naturalism and ascribes to us epistemic boundedness without considering a possibility that the core of the problem might lie in the fact that there is *more than one way we conceptualize our world*.

¹¹⁴ Ibid., p. 224.

¹¹⁵ Ibid., p. 225.

Davidson suggests that the realm of propositional attitudes and mental events in general is described by concepts which exhibit a logic completely different from the logic of the physical descriptions. We speak of beliefs, perceivings, desires etc. and this is very useful. In the spirit of Dennett's and Wegner's analyses we can understand these concepts as useful constructions and "user illusions" which enable us to function effectively in the world where we need to interact socially on an everyday basis. Wegner and Dennett showed that the self could be such a construction and a useful concept. That's why I believe that it is possible to infer that this concept belongs to the realm of the mental as described by Davidson.

The self, so to say, unifies different propositional attitudes and enables us to think of ourselves and others as persons – coherent agents with certain fixed characters and dispositions to act. There is no point in talking about propositional attitudes without reference to their origin – their "owner". As Davidson himself asserts, we always describe and explain mental events by referring to other mental events – this is the *holism of the mental realm*, and I see the concept of self as its crucial constituent, a "centre of its gravity". In order to properly explain a propositional attitude, we have to refer to other propositional attitudes belonging to the same agent, to the same *self*. The self is a part of an essential context which enables us to use our mental vocabulary meaningfully.

Davidson further suggests that the mental realm and mental events of which it consists escape the reach of purely physical explanation. This is, however, completely in accord with the view that all mental events are in fact physical. What is more, they are seen as physical within the boundaries of science which is attainable for humans. What is actually unattainable is the possibility to formulate strict laws connecting the physical and the mental *so described*. There isn't necessarily a gap in our knowledge of the natural reality; rather, we are beings who are capable of conceptualizing the world in different and mutually incompatible ways. We are capable of "having our cake and eating it too": we can see the whole world as completely physical and at the same time view it in terms of transcendent agents and their attributes – a view which deems the physical description irrelevant. We are a specific kind of natural dualists.

We are unable to "see" the self directly and explain it as consisting of certain "building blocks" arranged in such and such a way precisely because persons don't behave as bodies which can be described along the CALM terms. We are indeed a

part of the natural world, but we are so complex that only concepts obeying a different logic can provide us with the right means of interacting with each other. We get into trouble when we expect that the way we understand the physical world is a basic feature of our understanding as such. It is quite possible that the opposite is the case: that seeing the world in terms of persons is natural for us and that our sciences developed by transforming this understanding, just as Sellars suggests.¹¹⁶

The self sticks out as a perennial philosophical problem only when we fail to realise the fact of our “double life” - our dualism. On the one hand, we are “the folk”¹¹⁷ - social beings, free agents who interact with each other and understand themselves in terms of mental vocabulary; on the other hand, we are scientists (and philosophers) who parcel the world differently in order to uncover principles and laws of a different sort.¹¹⁸ The two “lives” cannot be mixed, and one cannot be understood in the terms of the other. We can, nevertheless, understand that we function this way and why this is the case. Both Davidson and Sellars show us that we live in two conceptual realms, and Wegner and Dennett come with a possible explanation for this: we live in conditions in which it is highly useful to view ourselves in a specific way - as conscious agents, selves, and persons. The reason why we infer our cognitive limits might not dwell in our inability to see the objective structure of the self but rather in our inability to see the two conceptual realms as one.¹¹⁹

McGinn certainly provides us with useful intuitions about the self: he correctly shows that it is not satisfying to accept reductionism or eliminativism, nor irreducibility or supernaturalist theories. The self must be natural because it is connected to other aspects of the world, but we have obvious troubles with accounting for its nature. Thus McGinn concludes that we are cognitively limited with respect to the real objective structure of the self.

On the other hand, there is a possibility to acknowledge our troubles with finding a proper place for the self in the natural world and at the same time know that

¹¹⁶ See Sellars, W., *Science, perception and reality*, p.10.

¹¹⁷ Referring to experimental philosophers’ term for laymen or nonphilosophers.

¹¹⁸ Ibid., p. 7.

¹¹⁹ Ibid., p. 39.

nothing escapes our cognitive capacities. If we follow the logic of Davidson's anomalous monism (and the "two images" of Sellars) we can come to understand this apparent paradox. When we talk about the self we use concepts that differ radically from those normally implemented in natural sciences. Wegner and Dennett provide an insight into the mechanisms that make this possible. We simply need to conceptualize complex beings such as ourselves in a specific and effective way, and thus we create the world of free responsible moral agents¹²⁰ who live their lives in intricate nets of propositional attitudes and mental events. We construct useful "user illusions" that are illusions only when we apply the scientific framework. Selves are real within the mental realm and manifest framework, and they enable us to interact with each other effectively. Scientific view would never enable us to understand these interactions, since they have their specific logic and rules.

There is no need to infer some objective real essence of the self. The self might be simply a crucial building block of the world built by our minds. This world exists parallel to the other world we constructed by yet different parcelling of the reality - the world of strict physical laws and postulated entities. We are nevertheless capable of transcending the two views (and probably there are many other such "sub-worlds") and reflect on the way we connect to the reality around us. Maybe such transcendence is the way we should at least attempt to follow whenever we have a feeling that we encounter our cognitive limits.

¹²⁰ Ibid., p. 38.

2 The Concept of Free Will¹²¹

The question of free will touches the first of the three essential sub-concepts of the self that we need to address in order to further articulate the specific conceptual area I have opened in the previous chapter. It is precisely the discussion over the possibility of compatibilism that reveals the tension arising from our natural tendency to see the world as full of persons on the one hand and attempts to explain everything using scientific approach on the other. What arises in this tension are the contours of the “double life” we have – our folk dualism.

Before I turn to concrete views, I would like to very briefly resume the basic possibilities with regard to the question of free will. *Determinism* plays the main role in the problem. It is a view asserting “*that every event has a cause. More precisely, for any event e, there will be some antecedent state of nature, N, and a law of nature, L, such that given L, N will be followed by e.*” Since this is supposed to be true of every event, it applies to the decisions and actions of human agents as well. Whatever the agent chooses, her decision is already fixed by the events that precede the decision.¹²²

One of the main reactions to this problem is *hard determinism* which accepts the truth of determinism and concludes that there is no space left for genuine freedom and responsibility. Then there is *soft determinism* or *compatibilism* that asserts “*that everything you should want from a notion of freedom is quite compatible with determinism.*” Despite being captured in the chain of causal necessity, “*it can often be true of you that you could have done otherwise if you had chosen,*” which means that you can keep a status of a responsible agent. The third

¹²¹ This chapter (except for the section “2.4 Support from experimental philosophy”) is mostly based on my published paper: Košová, M. (2015). Compatibilism and conscious will. *Filosofie dnes*, 7(1), 61-75. Available from <https://filosofiednes.ff.uhk.cz/index.php/hen/issue/view/15>

The study is set within the new context of the thesis; thus, some revisions were necessary. However, it would not be reader-friendly to cite the paper consistently throughout the chapter. To access the original paper, please visit the link cited above.

¹²² Blackburn, S. (2005). *The Oxford dictionary of philosophy* (2nd ed.). Oxford: Oxford University Press, p. 141.

possibility is to embrace *libertarianism* and claim that “*there is a more substantive, real notion of freedom that can yet be preserved in the face of determinism (or indeterminism).*” This can be connected to deconstruction of determinism, suggesting “*a special category of uncaused acts of volition*” or asserting “*that there are two independent but consistent ways of looking at an agent, the scientific and the humanistic,*” and that the apparent problem lies in their confusion.¹²³ The last libertarian response is akin to what I would like to suggest myself in what will follow, though I am more inclined to call this solution *compatibilism*.

In this chapter, I would like to introduce and discuss two approaches towards the problem of free will: Daniel Dennett’s compatibilism and Sam Harris’s hard determinism. Again, I don’t wish to provide an exhaustive overview of all possible views. I choose particular authors in order to point to those aspect of the problem that will serve me in my grasp of the main problem that is the focus of my thesis. My aim is to both problematize and find inspiration in ideas of these authors in order to better understand the confusion which comes together with the controversial question about the possibility of compatibility of free will and strict physical laws and the role that folk intuitions play in it. The main issue will be to characterize and confront two levels of thinking about free will – the unreflective intuitive level and the scientific, more fine-grained and reflective level. This brings once again the clash of the two Sellarsian images on the scene.

Parallel to the pervious chapter, in the first section of this chapter I will introduce Daniel Dennett’s conception, in the second Sam Harris’s criticism of Dennett and his own conclusions. In the third section I will try to point to the strengths and weaknesses of both approaches and use the strengths to reveal such approach towards the problem which would enable us to understand the confusion concerning free will. I will try to find a possible way to clarify it with the help of Sellars’s ideas about the two images. Finally, I will discuss the evidence from experimental philosophy and use it to demonstrate that the introduced theoretical solution makes sense in the face of empirical data. This all should lead me to a new formulation of compatibilism and an answer to the question how we can keep both intuitive and scientific understanding of free will while avoiding paradoxes and

¹²³ Ibid., pp. 141f.

tensions. Articulation of this solution should help me reveal further contours of folk dualism and develop in further detail what was sketched in the first chapter.

2.1 Apparent compatibilism

One of the crucial aspects of Daniel Dennett's conception is his attempt to demonstrate the illusoriness of our unreflective common-sense concept of free will which is closely connected to the concept of self. In order to understand the mechanism of how our concept of free will arises, it is important to turn to the concept of causation. Dennett mentions experiments conducted by Daniel Wegner which demonstrate the propensity of subjects to "*misattribute decisions to themselves that are in fact being made by somebody else.*"¹²⁴ Causation is a very problematic issue in the history of philosophy. We keep realising that the way we view causes and effects is largely dependent on our natural tendency to be "*overeager to interpret, to "notice" things causing other things when, in fact, both "cause" and "effect" are effects of complex machinery that is hidden from us – backstage, in effect.*"¹²⁵ It is all about *our* interpretation of what we observe. We are not getting the "real" causes served on a platter and, according to Wegner, our experience of conscious will arises from the process which interprets the connections between our thoughts and actions, not from the connections themselves.¹²⁶ When we feel that we do something consciously and voluntarily, it is only our interpretation of what is actually going on.

In fact, as Dennett points out, we only observe our decisions *arriving*; we never see the whole process of them being *made*. "*We have to see how we are going to decide something, and when we do decide, our decision bubbles up to consciousness from we know not where.*"¹²⁷ From this perspective it seems that we are strangely bereft of the real responsibility for our actions. We are only observers of the results coming from the impenetrable depths of our unconscious minds. This has a lot to do with "*the idea of the self as a unitary and cohering point of view on*

¹²⁴ Dennett, D. C., *Freedom evolves*, p. 243.

¹²⁵ Ibid., p. 244.

¹²⁶ Ibid.

¹²⁷ Dennett, D. C., *Elbow room: the varieties of free will worth wanting*, p. 78.

the world”, an illusion which arises when we are trying to come up with an interpretation of a certain action, and when we try to answer the question “*Did I do that?*”¹²⁸ This “illusion” is very useful since it helps us interpret in an effective way certain events we encounter. Under deeper observation, however, it doesn’t seem to make much sense.

By intuitively accepting the idea that our conscious self is really “the entity” that actually makes all the voluntary decisions, we implicitly accept some sort of “supernatural free will”. If we don’t assume that the decisions arrived to consciousness from intricate webs of unconscious processes, the only possible explanation is that they had to come from “nowhere”; they simply appear to conscious self somehow miraculously. Thus, conscious self is the only entity available to take the responsibility: “...*we exploit the cognitive vacuum, the gaps in our self-knowledge, by filling it with a rather magical and mysterious entity, the unmoved mover, the active self.*”¹²⁹ Dennett explicitly claims that free will thus understood does not exist. His position is succinctly captured in the following two sentences: “*If you are one of those who think that free will is only really free will if it springs from an immaterial soul that hovers happily in your brain, shooting arrows of decision into your motor cortex, then, given what you mean by free will, my view is that there is no free will at all. If, on the other hand, you think free will might be morally important without being supernatural, then my view is that free will is indeed real, but just not quite what you probably thought it was.*”¹³⁰ He suggests that free will is compatible with unconscious processes doing most of the work. We are not only the conscious tip of an iceberg; we are much more than that, all the intricate unconscious processes included. As Dennett often emphasises, there is a danger in excluding too much from our concept of self: “*As I never tire of pointing out, all the work done by the imagined homunculus in Cartesian Theater has to be broken up and distributed in space and time in the brain. It is once again time to repeat my ironic motto: If you make yourself really small, you can externalize virtually everything.*”¹³¹

¹²⁸ Ibid.

¹²⁹ Ibid., p. 79.

¹³⁰ Dennett, D. C., *Freedom evolves*, p. 223.

¹³¹ Ibid., p. 238.

In other words, according to Dennett, free will is real, but it needs to be redefined in the face of deeper insight into its underlying mechanisms, especially into the basis of the concept of self. On the one hand, he uncovers the illusoriness of the conscious self understood as the ultimate source of decisions and, on the other, he tries to save free will by widening the concept of self: we are not only the conscious observers but also the unconscious (and possibly deterministic) processes lying behind.¹³²

What is interesting, however, is that in Dennett's positive account the unreflective concept of "supernatural" conscious free will doesn't disappear altogether. This idea, as I see it, is very important for the constitution of what Dennett calls "*the atmosphere of free will*". This is an important conceptual atmosphere which enables us to think about the world in a certain way: we ascribe to people intentions, plans, hopes, etc., and we can honour them or blame them - all this because we perceive them as agents possessing free will: "*The idea that we have free will is another background condition for our whole way of thinking about our lives. We count on it; we count on people "having free will" the same way we count on them falling when pushed off cliffs and needing food and water to live...*"¹³³ It seems to me that the kind of free will that we ascribe to ourselves and to others has to be, in fact, "supernatural" in a certain way. We can be aware of its illusoriness, but we still employ it, intuitively and unreflectively, on the everyday practical basis. Dennett never explicitly states it like this, but I think that no concept of free will that entirely

¹³² This notion of free will is based on criticism of certain "overestimation" of the *causal* power of conscious will that appears also in Sven Walter's critic of Daniel Wegner. Walter draws attention to an unwarranted claim that people believe that their "feeling of conscious will" actually *causes* their actions. In his own words: "*Agency may be accompanied by the feeling of being the one who acts, and maybe even by the feeling that we cause our actions (although already that may be questionable), but actions are performed by agents, not caused. By acting, we can cause something, but to say that we cause our actions or cause ourselves to behave is (pace some die-hard agent causationists) at best misleading and at worst senseless. Agents don't cause what they do, they do it.*" (Walter, S. *Willusionism, epiphenomenalism, and the feeling of conscious will*, p. 2232.)

This point is in accord with the results of experimental studies exploring folk concept of free will, as I will show later in this chapter.

¹³³ Dennett, D. C., *Freedom evolves*, p. 10.

lacks certain “supernatural” aspect would really do the job. I will return to this problem later in this chapter.

The mentioned “atmosphere of free will” is something that had to come into being gradually. Dennett attempts to convey an exhaustive analysis of mechanisms which play crucial role in the process of constituting the atmosphere of free will. He starts by uncovering continual emergence of reasons, intentions and interests out of the complex set of conditions whose elementary roots are bereft of such attributions. *“In the beginning, there were no reasons; there were only causes. Nothing had a purpose, nothing had so much as a function; there was no teleology in the world at all. The explanation for this is simple: there was nothing that had interests.”*¹³⁴ The first “interests” evolved much later as a result of gradual process of complexity accumulation. They were not the full-fledged interests which are characteristic of human beings today, however. We only call them interests because we project concepts of our present perspective onto much simpler things which exhibit familiar patterns. In the spirit of this reflection we say that simple replicator’s interest is self-replication.¹³⁵ The interests became better defined with the development of the replicators’ abilities to *“defend their own interests”* and *“preserve this and that (their varieties of homeostasis)”*.¹³⁶ This means that certain concepts which we consider to be capturing some irreducible or “supernatural” aspects of reality are emergent – they are our cognitive reactions to the results of the initial conditions gradually changing in virtue of being formed under the influence of simple, arbitrary principles, namely the evolutionary principle as described by Darwin (his “strange inversion of reasoning”).

As I understand it, the fact that we tend to think in terms of reasons, intentions, etc., is due to *our* inability to view the whole complicated interplay of various factors. When we say that they “came to be”, we mean that *we* started to see the world this way. This could be nicely illustrated also by Dennett’s example describing two chess programs playing a match. Provided that the programs are complex enough, we view the combat as very suspenseful – from what we are actually able to consciously observe we can never predict which computer will win.

¹³⁴ Dennett, D. C., *Elbow room: the varieties of free will worth wanting*, p. 21.

¹³⁵ Ibid.

¹³⁶ Ibid., p. 22.

However, the programs are in fact determined: “*What from one vantage point appear to us to be two chess programs in suspenseful combat can be seen through the “microscope” (as we watch instructions and data streaming through the computer’s CPU) to be a single deterministic automaton unfolding in the only way it can, its jumps already predictable by examining the precise state of pseudo-random number generator. There are no real “forks” or branches in its future; all the “choices” made by A and B are already determined.*”¹³⁷ Unable to “see” the actual determinism of the situation, we are led to think of the computers as making choices and having numerous possibilities. Even though they are much simpler than human beings, we can still find it useful to treat them as agents and ascribe intentions to them.¹³⁸

According to Dennett, free will is not to be understood as some pre-existing feature of our existence; in fact, it evolves: “*It is an evolved creation of human activity and beliefs, and it is just as real as such other human creations as music and money. And even more valuable.*”¹³⁹ Concept of free will emerges as a result of our reaction to certain complex conditions. As was already suggested, intentions were not there at the very beginning. They came together with growing complexity, and so did the idea of responsible agents.

The atmosphere of free will consists of concepts such as “*intentional action, planning and hoping and promising – and blaming, resenting, punishing and honoring*”.¹⁴⁰ It seems that this conceptual complex is a basis for our understanding of us and other human beings as moral and responsible persons. Thus, as I understand Dennett, free will, personhood and moral responsibility go hand in hand, being connected through intentions, planning of acts, and subsequent praise or blame for these acts: as soon as beings evolve their own intentions and ability to plan acts following these intentions, they become responsible for their conduct because interaction between such beings naturally implies concurrent evolution of the atmosphere of free will and belief in free will (a kind of “bootstrapping”). We like or dislike certain actions, and it is only logical to trace them back to those “bundles of intentions” that are their source. Ascribing responsibility is a tool which helps us

¹³⁷ Dennett, D. C., *Freedom evolves*, p. 80.

¹³⁸ Ibid., p. 81.

¹³⁹ Ibid., p. 13.

¹⁴⁰ Ibid., p. 10.

influence behaviour of others as well as our own – praise encourages, blame discourages and the “air” of free will filled with responsibility makes us certainly think twice about our actions since they build our “moral image”.¹⁴¹

The idea of conscious free will is an emergent concept: even though it seems to capture something which escapes physical laws, it arose from completely natural conditions. It enables us to see the world in a specific way and helps us to act more effectively. Even though our perception of ourselves and other people is illusory on closer inspection, it is useful. This thought comes forward for example when Dennett discusses the shift from *design stance* to *intentional stance*. Imagine a designer who creates simple entities operating according to a set of simple principles. As he manages to develop more complex systems, there will come a point when he can start thinking of them as rational agents with intentions, beliefs, etc. This helps the designer to think of them on a higher level without being flooded by unnecessary details of the complicated mechanisms underlying the observable behaviour of the entities in question: “*It makes life blessedly easier for the high-level designer, just the way it makes life easier for us all to conceptualize our friends and neighbours (and enemies) as intentional systems.*”¹⁴²

The simplification brought about by the above described conceptualisation is similar to a familiar case of computer users that I mentioned already in the previous chapter. Dennett refers to the way software designers simplify and even distort the truth about the real workings of the computer so that it can be manipulated by the users intuitively. One can click and drag, hear various sound effects and orientate according to icons on the desktop – all this draws on usual and natural ways we perceive the world around us and act in it. Similarly, communication between people with “selves” provides access to such features of agents which are much easier to grasp and operate with than otherwise very intricate nets of brain processes. The concept of self certainly makes it easier to communicate what is going on in our brains and to influence other agents.¹⁴³ It is all about making things more effective on

¹⁴¹ As I will show in the following chapters of the thesis, this moral aspect is crucial for proper understanding of folk dualism.

¹⁴² Ibid., p. 45.

¹⁴³ Ibid., p. 248f.

a higher level of complexity. Reality may get distorted, but the most important thing is that the “trick” works in the end.

What we encounter here is the problem of confrontation of two different levels of our view of the world. As Dennett points out (referring to Sellars), on the one hand, there is an unreflective every-day view - all the phenomena we see with naked eye, middle-sized objects, and rates of change, etc. (the *manifest image*). On the other hand, we also have the *scientific image*. For example, “we understand that while “water” is a mass noun for us, water is also a swarm of countable molecules, whose trajectories are trackable in principle, and sometimes even in practice (with the aid of prosthetic extensions of our senses).”¹⁴⁴ We are capable of abandoning our manifest image and start looking at the world differently, using “*more fine-grained level of description*”. While adopting this new outlook, we start realising how problematic our everyday common-sense thinking really is. However, the truth is that we need our intuitive understanding which provides the only way to think on a day-to-day basis. We can clearly see it in the case of the free will problem: we are deliberators, and if we want to deliberate effectively, we must be faithful to certain unreflective concepts (e.g. “open future”, even if determinism is true).¹⁴⁵

I believe that effectiveness is really the key issue here: we have to work with what we have, and we have finite and bounded epistemic equipment. If we were able to process all the information concerning the intricate causal interactions pertaining to micro-level in short time, we wouldn’t need “user illusions”. We do need the concept of self precisely because it provides us with the ability to predict behaviour and orientate in the world of complex beings while we only have to process relatively small amount of information. We project “intentions” and “beliefs” into others and ourselves because it works in the end – we are capable of accurate predictions and effective interactions in the Davidsonian mental realm. Concepts concerning personhood serve us well and thus influence dramatically the way we intuitively view reality – even though in this case we are not “knowers”, as we tend to think, but in certain sense “constructors” of useful conceptual tools. As by using computers, we don’t learn about the mechanisms which make all the applications possible; we simply use them and learn on this user-level.

¹⁴⁴ Dennett, D. C., *Elbow room: the varieties of free will worth wanting*, p. 114.

¹⁴⁵ *Ibid.*, p. 114f.

In my opinion, the main advantage of Dennett's conception of free will is that it brings forward a certain idea of "concept-emergence". A level of reality which seems utterly unique and irreducible (as the above discussed level of the atmosphere of free will) can arise from much simpler elements, provided that these elements are arranged into sufficiently complex structures (beings like us) entering mutual interaction. The important thing to realise is that the feeling of novelty and irreducibility results from the way the complex situation in question is perceived by us. Various levels of thinking about the world are to be considered. On the practical or day-to-day level it would be highly ineffective to try to keep track of all the details, and that's why appropriate simplification or even distortion comes in handy. In other words, intricate complexity yields novel properties *for us*, because *we*, finite and cognitively bounded beings, need to effectively handle this complexity.

From the everyday perspective, the scientific image is unnatural. Our basic understanding of the world is based on various intuitions and unreflective folk concepts which work like useful shortcuts or user illusions. Once we look deeper into the mechanisms that ground the familiar phenomena, we suddenly encounter completely different world which so often contradicts our everyday perception. Thus, in order to avoid confusion, it is important to realise that the everyday view and the scientific view have to be carefully distinguished. Dennett goes in this direction, but I don't think that he fully articulates the role of the intuitive level. He redefines free will in terms of the scientific view and by this introduces certain confusion into the problem – the "redefined free will" belongs to the scientific framework, but the original "intuitive free will" is something quite different. In my view Dennett confuses the two possible meanings and bases his compatibilism on this confusion.

2.2 Changing the subject

Sam Harris criticises Dennett's conception as problematic in his own account of free will whose style is more accessible to wider public and not as philosophically deep as Dennett's account, but it rightly points to some important issues, especially because it reflects upon folk beliefs. This has to do with the role that folk intuition plays in the problem. Harris emphasizes the subjective strength of our everyday common-sense concept of free will. He directly opposes Dennett's compatibilism, especially his claim that we are not only the conscious tips of an iceberg but also the

intricate unconscious neural processes. Harris stresses repeatedly that the free will problem is based on psychological fact – the feeling most people have about their free will: “*Compatibilists generally claim that a person is free as long as he is free from any outer or inner compulsions that would prevent him from acting on his actual desires and intentions. (...) The truth, however, is that people claim greater autonomy than this. Our moral intuitions and sense of personal agency are anchored to a felt sense that we are the conscious source of our thoughts and actions.*”¹⁴⁶ According to Harris, compatibilists simply “change the subject”: they ignore the subjective feeling people have about their status as conscious agents and serve us with a specific concept of person instead. What makes the problem of free will so acute is the *feeling of agency and moral responsibility*, and to ignore this is to miss the whole point.¹⁴⁷ According to Harris, there is no place for compatibilism because it ignores the only kind of free will worth talking about – and this kind of free will simply doesn’t exist.

To support his claim that we don’t have free will, Harris too turns to the problem of the attribution of agency and mentions the scientifically described cases showing the unreliability of our interpretation skills. According to him, our interpretation of the cause-effect relation between our thoughts and actions is even more erroneous than we tend to think: “*There is no question that our attribution of agency can be gravely in error. I am arguing that it always is.*”¹⁴⁸ It is not only actions but also intentions whose origins we interpret incorrectly. The problem is that we are consciously aware of our intentions, and we intuitively believe these intentions to originate from our conscious selves. Their true source is, however, hidden from us and belongs to the realm of unconscious brain events that we don’t intend.¹⁴⁹

For Harris it is all about the subjective feeling of agency and our intuitions. The concept of free will he is interested in is the unreflective concept which, according to him, most people share. This is what Dennett would call “supernatural” free will, since under closer observation it doesn’t make sense. It implies decisions

¹⁴⁶ Harris, S. (2012). *Free will*. New York: Free Press, p. 27.

¹⁴⁷ Ibid., p. 31.

¹⁴⁸ Ibid., p. 31.

¹⁴⁹ Ibid., p. 32.

coming from nowhere and suddenly appearing in our consciousness. If the conscious self was their ultimate author (providing the unity that McGinn talks about in his account of the self), we would stand face to face with a strange idea that we create ourselves *ex nihilo* – that our decisions - building blocks of our moral character, pop up suddenly without any prior warning thanks to the god-like power of our conscious self and *nothing else*. However, when we give this problem a deeper thought it seems only natural that every decision has to be based on something. We need some prior background of knowledge and experience in order to deliberate, and it should not be very surprising that *unconscious* brain processes play the main role. When we really think about it we realise that there is no reason to presuppose an agent independent of all the possible influences. On the contrary, both outer influences and those coming from our own brain are in fact necessary for a process of deliberation and the resulting decision to take place. Harris agrees with all this, but for him it doesn't imply that we are more than the conscious self. Rather, it implies that we, conscious selves, live in an illusion.

Harris, unlike Dennett, doesn't try to "save" free will; he simply states that we don't have it. He "analyses free will away" and doesn't seem to find any particularly positive role for the unreflective concept in our everyday lives. He acknowledges that thinking about free will in terms of its illusoriness might have some bad impact on certain moral tendencies (he mentions example of students who cheated more after being confronted with an argument against the existence of free will). On the other hand, he claims that "the truth" could possibly increase one's "*feelings of compassion and forgiveness*".¹⁵⁰ Anyway, his ultimate claim reads: "*The illusion of free will is itself an illusion.*" We might feel that we are the conscious authors of our decisions, but as soon as we try harder and explore our experience more thoroughly, we realise that we don't even feel supernaturally free anymore: "*It is not that free will is simply an illusion – our experience is not merely delivering a distorted view of reality. Rather, we are mistaken about our experience. (...) Our sense of our own freedom results from our not paying attention to what it is like to be us. The moment we pay attention, it is possible to see that free will is nowhere to be found, and our experience is perfectly compatible with this truth. Thoughts and*

¹⁵⁰ Ibid., p. 44.

intentions simply arise in the mind. What else could they do?”¹⁵¹ There simply seems to be no place left for conscious free will.

2.3 Free will and the two conceptual realms

In my opinion, both Dennett’s and Harris’s conceptions share a common problem: neither Dennett nor Harris distinguishes properly between the two levels of conceptualization of free will.

Dennett refuses the unreflective concept (the one emphasized by Harris) as “supernatural” and illusory and redefines free will in terms of broader understanding of its mechanisms. On the other hand, he seems to acknowledge the importance of our thinking about ourselves and others as possessing free will. However, my problem with his conception is that when he talks about the atmosphere of free will he seems to be talking about the unreflective view of free will (perhaps the conscious free will), but he doesn’t explicitly acknowledge its positive role. The question is whether the atmosphere of free will would be preserved provided that we forgot about the conscious “homunculus agent” or any “supernatural” concept altogether and thought about our own free will scientifically. I am not of that opinion. If we didn’t use the “shortcut” or “user illusion”, we would not be able to act effectively, “under the idea of freedom”. It is quite possible that we have to view ourselves and others as conscious agents escaping physical causality in order to preserve the effective functioning of our moral interactions.¹⁵² This does not mean that we cannot also be aware of the true mechanisms, but we simply don’t embrace this scientific view when we are acting in our social world on everyday basis. I appreciate

¹⁵¹ Ibid., p. 56. This might seem paradoxical and lead us to think that Harris contradicts himself. I believe, however, that he simply tries to show precisely that the paradox is a crucial feature of our understanding of free will. Conscious free will is an unreflective intuition - it is like a fuzzy picture: it makes perfect sense until we look at it too closely. It is simply not meant to be looked at too closely.

¹⁵² See e.g. Baumeister, R., Maslach, E. J. C., DeWall, C. N. (2009). Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin*, 35(2), 260-268. The authors suggest that their experiments point to a possibility that “*disbelief in free will increases aggression and reduces helpfulness*”.

Dennett's attempt to redefine free will so that we incorporate also unconscious processes into our concept of self, but I agree with Harris that this is simply changing the subject. Dennett claims that we have free will, but *this* free will is not *the* free will whose concept actually enables us to enter the "atmosphere of free will".

I believe that also Harris misses an important point by not paying enough attention to the role of our intuitive concept of free will: it really seems that people normally act "under the idea of freedom" which means that they think of themselves and of others as free conscious agents who escape deterministic laws. The fact that this idea doesn't make sense under closer inspection has nothing to do with the way we function in the world on a day-to-day basis. There is a level of conceptualisation which plays crucial role for us as complex beings interacting with other complex beings. We need shortcuts and "user illusions" to get by in this environment – we need to be "dualists". We can also realise, by using a microscope, that water isn't really what we normally see it to be. But this doesn't mean that we *should* deny its liquidity and transparency without further qualifications.

I have to agree with Harris's criticism of Dennett's "changing the subject" - it seems to me that he confuses the issue by redefining free will and putting the unreflective concept of conscious free will aside. However, Dennett does a good job by introducing different possible ways of understanding complex phenomena (e.g. our switching to "intentional stance"). If we follow Harris, it may lead us on our way to acknowledge the subjective strength of the unreflective folk concept. Following Dennett, in turn, can enable us to find a proper place for *this* free will. Combining the two, we do justice to our common-sense concept, and at the same time we save its validity by looking at free will while distinguishing different possible stances. We get back to Davidson and his anomalous monism: we have to distinguish between the two realms and realise that the mental realm is not reducible to the physical realm. Each of these two realms has its own logic.¹⁵³

In order to illustrate the point, I would like to return to the parallel between our concept of free will and that of water. Once again, unreflective understanding of free will implies the idea that there is some ultimate unit – the conscious self who is the sole author of all the free decisions and who transcends physical laws of

¹⁵³ I will refer to how exactly free will fits into the mental realm in the next section – "Support form experimental philosophy".

causation. This is the homunculus, the explanatory entity of the first order – something that makes the mechanism of free will seem “supernatural” because we simply have to take it for granted and cannot ask about further, more fine-grained explanations which would be more akin to scientific inquiry. When we analyse the phenomenon closer, however, we discover immense net of unconscious processes and failures in our ability to reveal the actual causal relationship between our thoughts and actions. From scientific perspective there is no conscious self which would be able to escape physical laws. In my opinion, something similar happens when we think about water. On a day-to-day basis water is something liquid, transparent, something we can drink and something we can drown in, etc. From scientific perspective (the scientific image) water is neither liquid nor transparent – it is a collection of molecules of H_2O . This is supposed to be the true nature of water, viewed as a result of more fine-grained analysis. On this level of understanding it doesn’t make sense anymore to talk about manifest properties of water as they appear to our senses since this would be just a coarse approximation or even distortion.

We can, however, use the scientific view to explain the effects which appear in our unreflective and “crude” level of reality. We explain liquidity by referring to behaviour of the H_2O molecules, for example. In the case of free will it is not so much different. We discover the mechanisms which elicit in us the specific subjective feelings of conscious agency. However, this doesn’t mean that we should deny the validity of the unreflective intuitive thinking. Both the concept of liquidity and the concept of conscious free will are emergent: they represent the phenomena in question in such a way that many of the “scientific details” can be abandoned so that we can orientate more effectively in our world. They are clever shortcuts which may distort the actual mechanisms standing behind the phenomena, but which are very useful in day-to-day practice, nonetheless.

In the case of free will we don’t see the whole causal chain of brain processes leading to our decisions in the same way we see how liquidity arises on a basis of molecular structure etc. (the case of water seems more intuitive, so to say). However, the main difference between the two examples dwells in the fact that science doesn’t deny liquidity in the same way it seems to deny free will: in the case of free will the denial of this phenomenon from the position of the scientific framework seems so much more prominent than in the case of water because it has an immense impact on

how we understand ourselves as moral beings. The fact that we don't see molecules with the naked eye doesn't really matter so much, and we feel free to let it pass.

The point I am trying to reach is well illustrated not only by the previously mentioned anomalous monism of Davidson, but also by what Sellars says about the conflict between the everyday and the scientific framework:¹⁵⁴ *"...the claim that physical objects do not really have perceptible qualities is not analogous to the claim that something generally believed to be true about a certain kind of thing is actually false. It is not denial of a belief within a framework, but a challenge to the framework. It is a claim that although the framework of perceptible objects, the manifest framework of everyday life, is adequate for the everyday purposes of life, it is ultimately inadequate and should not be accepted as an account of what there is all things considered."*¹⁵⁵ Only when we approach free will problem from this perspective can we come to understand why "the truth" seems so surprising, unintuitive and controversial.

The core of the problem seems to be this: confusion happens when we mix the frameworks and the realms. I believe that it is simply impossible to build a conceptual bridge which would smoothly connect scientific and intuitive accounts of free will. We encounter an unavoidable abyss here. Scientific thinking is based on carefully articulated explanations and fine-grained analyses. The common-sense concept of free will ends up being revealed as unsatisfactory because of the weakness of its explanatory power. Wegner points out that the unreflective idea of free will is "homunculus-based". We postulate a homunculus who decides things without any prior causes which would have some impact on the decisions. We come to *"an explanatory entity of the first order. Such an explanatory entity may explain lots of things, but nothing explains it. (...) A first-order explanation is a stopper that trumps any other explanation, but that still may not explain anything in a predictive sense."*

¹⁵⁴ Both authors were inspired by Kant, who brought this idea of the "clash between the images" on the scene long before them. For Kant, *"while the empirical or phenomenal self is determined and not free, the noumenal or rational self is capable of rational, free action. But since the noumenal self exists outside the categories of space and time, this freedom seems to be of doubtful value."* (Blackburn, S., *The Oxford dictionary of philosophy* (2nd ed.), p. 141.)

¹⁵⁵ Sellars, W., *Science, perception and reality*, p. 27.

(...) *There cannot be a science of this.*”¹⁵⁶ We may intuitively feel that there is such a homunculus, but as soon as we assume scientific approach and analyse the deliberation process more precisely, we uncover the pitfalls. This is the moment when we, together with Harris, might want to say that there simply is no such thing as free will.

Harris makes the above-mentioned unwarranted step: he mixes the two frameworks. He adopts scientific framework and tries to implant the fine-grained concept of free will to the framework of our common-sense folk intuitions. This results in confusion and tension because, scientifically speaking, conscious free will is a nonsense, and, “naturally” or “intuitively” speaking, we don’t like the idea of determinism applied to ourselves as beings living in the world with moral dimensions. Dennett does distinguish between the frameworks – he is aware of the fact that on a certain level of complexity it is useful to conceptualise things differently (a very good example is his “intentional stance”). But he doesn’t state it clearly that it is the “supernatural” conscious free will which serves us so effectively on the level of interactions between moral agents. He changes the concept of self to save free will, and by this he smuggles in the scientific framework and the physical realm. His “atmosphere of free will”, I believe, has to belong to the manifest, common-sense framework, and the mental realm. This means that he shouldn’t say that supernatural free will doesn’t exist. He should say rather that it exists within its own manifest framework and mental realm and is illusory within the other – the scientific framework and physical realm.

Yet another thing which requires clarifying is the comparative adequacy of the frameworks. Even though we tend to perceive the scientific framework as the “true” one, or the one we should prefer (Sellars’s “scientia mensura”), this could be in fact misleading. Science can indeed provide us with fine-grained detailed explanations by uncovering the intricate mechanisms behind things, but there is a functioning way of understanding the world which *ignores* these details. It is this very fact of simplification which gives rise to new and wonderful level of existence – a world of incredibly complex and yet epistemologically bounded creatures. This paradoxical combination of complexity and epistemic boundedness is the true soil where free will can flourish. To handle their own complexity, the creatures need to

¹⁵⁶ Wegner, D. M., *Who is the controller of controlled processes?*, p. 20.

learn to understand¹⁵⁷ themselves in a specific way: being finite and unable to process all the possible information available, they have to think in shortcuts and clever simplifications. Their world is a world where the intuitive unreflective concepts are its real building blocks. The world of conscious agents is real on its own level and cannot be torn down by scientific analysis. Science can only describe how various building blocks came into being, but the bricks and pillars of the manifest image still work together to give rise to a coherent building. In other words, mental realm keeps its holism.

I believe that it is also very important to stress how the two frameworks differ in the way we employ them. Scientific stance can be adopted only temporarily and under specific conditions: it can take us considerable amount of time to analyse a certain phenomenon, and this analysis has to deal with such complicated information that we have to exert considerable amount of effort to grasp it. This all, of course, is in contrast with prompt interpretation of our environment facilitated by our intuitions.¹⁵⁸ Despite scientific thinking brings quality and reflexion into our lives, in many cases it can never become our day-to-day mode of orientation in the world – it is unnatural and ineffective. Manifest framework, by having its own building blocks (unreflective concepts which work as useful simplifications) and its own rules, has its own indisputable validity. It is a world on its own.

By keeping the two frameworks apart we can better understand the specific status of conscious free will. We can see its illusoriness from one perspective and acknowledge its validity from the other. If we adopt the scientific perspective and try to suppress our intuitions for a moment, we can gain an understanding of how our worldview works on various levels. We can get insight into the process of gradual coming to be of intentions, purposes, and even free will. We must remember, however, that we are still those same finite and epistemologically bounded creatures who need to function effectively in an immensely intricate reality. We simply cannot get by without useful simplifying concepts, because they make us understand the world on one very important level – a level we cannot escape because it is always

¹⁵⁷ This kind of understanding is not really meant to simply help us explain behaviour of complex beings; rather, it is meant to provide us with a tool for *interacting* with them.

¹⁵⁸ This is in accord with the general idea behind dual process theories.

with us, wired in our brains. The only way to avoid confusion is to fully realise that we are dualists in an important sense: we are capable of adopting two different perspectives and two modes of understanding reality whose particular roles have to be carefully distinguished.

2.4 Support from experimental philosophy

In the previous paragraphs, we have addressed different theoretical approaches towards the problem of free will. Since the debate touches concepts that are ascribed to people in general, it is apposite to look for an actual evidence that these beliefs are present in laymen. These findings will help us both defend and criticise the theoretical outlooks presented in the previous sections and thus better define the free-will aspect of folk dualism.

While looking at the views of Dennett and Harris, we have encountered the concept of *conscious will* that brings the crucial “unity” of the self into the picture and found it useful in speculations about the “natural” concept of free will – the free will people actually care about. The role of consciousness in the question of free will seems to be crucial or even self-evident in many different notions of free will, but the details about this role remain unclear. Nahmias, Allen, and Loveall point to the absence of the thematization of this problematics in the existing literature.¹⁵⁹ One of the possibilities why consciousness seems to be crucial for free will is the suggestion that, in one of the libertarian interpretations, “*conscious self can be an uncaused caused, free from deterministic chain of cause and effect in the physical world.*”¹⁶⁰ I will return to the role of consciousness in the notion of free will in the fourth chapter. Now I would like to focus on the notion of conscious will as an *uncaused cause* or an entity that *escapes causal necessity of the physical world*, because it seems to be an important aspect of the folk notion of free will, as I will show in the following paragraphs.

¹⁵⁹ Nahmias, E., Allen, C. H., & Loveall, B. (2020). When do robots have free will? Exploring the relationships between (attributions of) consciousness and free will. In B. Feltz, M. Missal & A. Sims, (Eds.), *Free will, causality, and neuroscience* (pp. 57-80). Brill. Retrieved from <https://brill.com/view/title/38676>

¹⁶⁰ Ibid., pp. 61f.

Experimental philosopher Joshua Knobe considered the volume of experimental results from the studies focusing on the folk notion of free will and came up with interpretation that points to what he calls the “*transcendence vision*” – in short, people conceptualize human actions in a way that transcends pure scientific vision.¹⁶¹

In the cross-cultural study by Sarkissian et al.,¹⁶² people from various different cultures (United States, Hong Kong, India, and Colombia) tended to claim that in our universe human action escapes causal determinism. Knobe’s *transcendence vision* seems to be a plausible explanation of these experimental findings.¹⁶³ Things get more complicated with studies targeted at the relationship between determinism and moral responsibility. When people were asked about cases when the agent was causally determined (e.g. via deterministic universe scenario) to perform certain abstract action, they tended to conclude that the agent was not morally responsible for this action. However, when the action was described concretely and was particularly violent and morally reprehensible, people considered the agent morally responsible despite the suggested determinism. The reason for this pattern of answers is still not clear, but Knobe uses these findings to ask a different question: why causal determinism seems to be a problem for moral responsibility at all (as in the abstract cases)?¹⁶⁴

Knobe refers to a series of studies by Nahmias and Murray¹⁶⁵ that might be the key to solving the question. Participants in the studies tended to agree that in deterministic universe, person’s beliefs and desires have no effect on their actions.¹⁶⁶ Knobe suggests that “*people are conceptualizing the relationship between an agent’s beliefs and desires and his or her actions in a way that is radically different from the*

¹⁶¹ Knobe, J. (2014). *Free will and the scientific vision*. In E. Machery, E. O’Neill (Eds.), *Current controversies in experimental philosophy* (pp. 69-85). Abingdon: Routledge.

¹⁶² Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., Sirker, S. (2010). Is belief in free will a cultural universal? *Mind & Language*, 25(3), 346-358.

¹⁶³ Knobe, J., *Free will and the scientific vision*, p. 73.

¹⁶⁴ Ibid., pp. 73ff.

¹⁶⁵ Nahmias, E., & Murray, D. (2010). Experimental philosophy on free will: An error theory for incompatibilist intuitions. In J. Aguilar, A. Buckareff & K. Frankish (Eds.), *New waves in philosophy of action* (pp. 189-216). Hampshire, England: Palgrave-Macmillan.

¹⁶⁶ Knobe, J., *Free will and the scientific vision*, p. 76.

way they would normally conceptualize the relationship between a fire and the destruction of the house.”¹⁶⁷ In other words, typical cases of physical causation are not analogous with the mechanisms of human action. Human moral agents escape the realm of physical causation, which is in accord with the “conscious will” notion I touched in the previous sections.

Now we are getting close to a model that bears connection to Davidson’s idea of the holism of the mental realm. Knobe points to the fact that people often use language of reasons when explaining agent’s actions – they refer to beliefs and desires of the person in question. The most important point is that the causation that plays a role here is not compatible with the causation as we normally understand it. Instead of saying that “the action was caused by a belief” people tend to say that the action was “chosen for a reason.” This is in accord with Knobe’s *transcendence vision*. When people talk about an action in the language of reasons, they refuse the possibility that the behaviour of the agent in question is causally determined. This is exactly the explanation that the experimental studies point to, since in the case of deterministic universe the respondents refuse the effect of beliefs and desires on the agent’s actions, because they probably refuse the possibility of an agent acting for reasons in this type of universe.¹⁶⁸ By his concept of *transcendence vision* Knobe simply refers to “the idea that human actions are radically different from other sorts of events.”¹⁶⁹

Based on these and other similar findings from studies focusing on many different concepts (e.g. happiness, knowledge, or causation), Knobe arrives at a conclusion that the way folk understand the world differs quite radically from understanding typical for sciences that is often ascribed to them.¹⁷⁰ As for his *transcendence vision*, it is perfectly compatible with my suggestion that Davidson’s idea of anomalous monism could help us understand the special status of the concept of free will. Folk apparently believe that free will somehow transcends the rules of the physical world (as Harris suggested), and by moving free will into the

¹⁶⁷ Ibid., pp. 76f.

¹⁶⁸ Ibid., pp. 77f.

¹⁶⁹ Ibid., p. 79.

¹⁷⁰ Ibid., p. 82. This is also in accord with the view concerning the folk concept of consciousness to which I will refer in the fourth chapter.

Davidsonian mental realm of reasons we can have our own kind of compatibilism in which free will is not a mere illusion, but an important member of the atmosphere that enables us to conceptualize persons.¹⁷¹

Even though Knobe's account brings satisfactory interpretation of the experimental results, we should look closer at the way folk intuitions escape pure scientific vision. As Eddy Nahmias and Morgan Thompson suggest, Knobe's account might be too simplistic.¹⁷²

Both Harris and Knobe present folk view of free will as incompatibilist because their "transcendent self" is supposed to be escaping the whole causal order. Nahmias and Thompson suggest, inspired by Harris, that this should also apply to brain-imaging technology. If we were able to see the whole process of decision-making and were able to tell how the subject will decide even before the information arrives to their consciousness, this would rule out free will in the eyes of most people. Their experimental results, however, deny this prediction. Most of their respondents agreed that such technology is possible in principle and that it would not rule out free will and responsibility.¹⁷³

Nahmias and Thompson thus refuse Knobe's *transcendence vision* and suggest that so-called "*naturalistic vision*" reflects the folk view of free will much better. They understand this alternative view as a middle position between *scientific vision* and *transcendence vision* because they believe that "*ordinary people have a fuzzy understanding of the nature of mind, action and free will*" and that the folk "*theoretical commitments are relatively noncommittal and revisable.*" Naturalism is a view that understands free will as a phenomenon that doesn't transcend natural laws

¹⁷¹ "But the explanations of mental events in which we are typically interested relate them to other mental events and conditions. We explain a man's free actions, for example, by appeal to his desires, habits, knowledge and perceptions. Such accounts of intentional behaviour operate in a conceptual framework removed from the direct reach of physical law by describing both cause and effect, reason and action, as aspects of a portrait of a human agent. The anomalism of the mental is thus a necessary condition for viewing action as autonomous." (Davidson, D., *Essays on actions and events*, p. 225.)

¹⁷² Nahmias, E., & Thompson, M., *A naturalistic vision of free will*.

¹⁷³ Ibid., p. 86.

and is at the same time open to compatibilist solutions, such as those characteristic of folk view, as Nahmias and Thompson interpret it.¹⁷⁴

Knobe is right in his claim that most ordinary people have a concept of free will that is structurally different from typical scientific theories. Nahmias and Thomson see his mistake in introducing the idea of *transcendence vision* – a kind of theory of its own. Instead, they propose that the folk are “*theory-lite*”. They simply don't build their concept of free will around a robust theory of any kind. They understand actions of a human agent via the language of reasons, but they don't commit to any kind of underlying metaphysical theory about the nature of the mind and its processes.¹⁷⁵

Even if we take into consideration dualistic tendencies of people, such as their belief in souls¹⁷⁶, we cannot find any kind of well-established substance dualism behind them. Folk beliefs are vague and open to concrete details of how the “soul” fits in the world. The folk concept of “soul” is a kind of placeholder for “*whatever underlies the set of capacities humans have for thinking, feeling, and acting*” and a specific uniqueness of a person. The concrete realisation of this placeholder remains very vague. What is more, various experiments show that the vagueness goes as deep as to enable a concept of free will that survives even if the idea of an immaterial soul falls out of the picture.¹⁷⁷

Nahmias and Thompson demonstrate the strength of their account by further experimental results. Their survey with over 100 American university students was based on a scenario about a brain-scanner technology capable of using information about brain activity to predict with 100% accuracy all person's thoughts and decisions even before the conscious awareness of them kicks in. Most of their participants agreed that if such a technology really existed, the person in question and people in general would still have free will and be responsible for their decisions.¹⁷⁸ Nahmias and Thompson believe that their respondents, contrary to

¹⁷⁴ Ibid., p. 88.

¹⁷⁵ Ibid., p. 89.

¹⁷⁶ I will address the topic of the concrete content of the folk concept of soul in the third chapter.

¹⁷⁷ Ibid., pp. 90f.

¹⁷⁸ Ibid., pp. 91ff.

Knobe's theory, don't see any serious conflict in the fact that decisions are predicted and caused by both brain states and reasons, thus they are probably *theory-lite* with regard to the way mental states and brain states relate to each other. Free will and human reasons survive in the world of brain-state causality.¹⁷⁹

Even though the results and their interpretation are in certain tension with Knobe's account, I believe that Nahmias and Thompson haven't succeeded in addressing it properly. Causality in the realm of brain activity might not be viewed by the respondents as an example of purely physical causation, as Knobe uses this concept. Brain activity concerns mental realm, and the participants might have understood it precisely along these lines. What is more, when Nahmias and Thompson altered the scenario and included manipulation of the brain activity by neuroscientists, most people refused to view this situation as compatible with free will and responsibility.¹⁸⁰ I believe that the scenario about the neuroscientist causing someone to decide in a certain way captures the idea of physical causality much more aptly, and it is possible that the respondents' change in the pattern of answers is a response to this fact. Anyway, I still agree with Nahmias and Thompson that folk concepts are theory-lite and open to many different metaphysical details.

The authors further propose a clever principle that enables certain compatibility between their and Knobe's theories. "*Causal competition*" principle predicts the following: "*People will be reluctant to hold an agent responsible for behavior when they interpret his or her behavior as being fully caused by factors that do not include any of his or her reasons (or by processes that do not include any of the agent's reasoning).*"¹⁸¹ The difference lies in the way people interpret different scenarios. Nahmias and Thompson simply suggest that thanks to their "theory-lite approach", people view free will and responsibility as compatible with a wider range of scientifically coloured scenarios than Knobe is willing to admit. Above all they refuse to ascribe to the folk the idea that free will is dependent on an agent capable of deciding "from nowhere", as Harris seems to claim. Folk are willing to accept certain scientific explanations of our actions without losing their beliefs about free will as long as they enable the language of reasons to survive. Anyway, I believe that it is

¹⁷⁹ Ibid., p. 93.

¹⁸⁰ Ibid., pp. 94f.

¹⁸¹ Ibid., p. 97.

necessary to stress that the “scientific” cases that seem to be compatible with free will in the eyes of laymen are very specific: they concern brain states, and these might be viewed as connected to mental realm rather than to the physical realm and the typical cases of causation.

As far as I am concerned, the experimental results described in this section support my idea of compatibilism – the need to distinguish between the two conceptual realms. People seem to have a concept of free will that transcends purely scientific vision. The way people conceptualize human action is based on the language of reasons that strongly refers to Davidson’s mental realm. At the same time, we don’t have to claim that the folk have perfectly articulated theory of free will. Instead, I agree with Nahmias and Thompson and suggest that the folk concept of free will is *theory-lite*. This is in accord with the previous idea that people use “user-illusions” and shortcuts in order to be able to view persons effectively instead of deep and elaborate theories that would yield a perfect understanding of the phenomenon. This means that the folk still might have certain unreflected idea of conscious will as a transcendent “unmoved mover” of the mental realm that escapes physical causation, but the details of this notion remain unclear due to the fuzzy nature of folk theories.

I have attempted to show that both Dennett and Harris approach the problem of free will in a way that is in conflict with our intuitive understanding of ourselves and other people. They simply state that conscious self is not the ultimate author of our decisions. Harris sticks to determinism, while Dennett tries to save free will and comes up with a version of compatibilism. However, I believe that this is not the real compatibilism. Harris correctly criticises Dennett for his “trick” – redefining free will so that it fits scientific framework and then claiming that *this* “Dennettian” free will exists, and thus compatibilism is true. The real compatibilism would have to acknowledge, just as Harris points out, our subjective feeling of conscious agency as the free will we in fact care about. In other words, certain transcendence steps into the question, even though this transcendence is not based on any exhaustive metaphysical theory. Empirical data from studies of folk intuitions point precisely in this direction.

I proposed a way how to get to a compatibilism that is in better accord with human intuitions. It is inspired by Dennett’s theory of different possible “stances”,

Sellars's theory of different frameworks or images and Davidson's theory of the two realms. Certain way of understanding the world is natural for us – we are wired this way, and this fact enables us to effectively handle environment where complexity reaches new levels. On the other hand, we are also capable of more fine-grained approach – we became sophisticated enough to perform deeper analyses of the phenomena we encounter in our everyday lives. This led us to many unintuitive conclusions. The problem is that we tend to mix the frameworks. The only way we can clarify the situation is to acknowledge the abyss between the two frameworks. To say that free will doesn't exist is like to say that physical objects don't have perceptible qualities. They indeed do have perceptible qualities – for *us* on the everyday level and with our theory-lite concepts. We indeed do have free will – it is real *subjectively*, for *us* as finite beings who try to make their way through the world which is too complicated to be perceived in detail. Manifest image has its own validity; it is a world on its own which works in its specific way.

The folk concept of free will that arises from the manifest image is a good illustration of the “folk dualism” – the view that sees persons as transcendent in relation to the world described by science. As I already mentioned in the previous chapter, *causal agent* seems to be one of the building blocks of the folk concept of person. In the next two chapters I will continue to fill the mosaic by focusing on the concept of *essential self* (chapter 3) that points to the problem of personal identity and belief in souls, and the concept of *subject* (chapter 4) that opens the problem of consciousness and also returns to the question of the role of consciousness in the folk concept of free will.

3 The Concept of Soul and the Essential Moral Self

I have already touched the concept of self as such and turned to one of its crucial aspects – the concept of autonomous agent. At this point it is time to address the other two concepts that belong to the concept of “minimal self” – the *essential self* and the *subject*.

In this chapter I will start with the concept of *essential self* and its relation to the folk concept of *soul* and *personal identity*. In the previous chapters I have pointed to the basic “clash between the images”. We already know that one of the things that make persons irreducible in respect to scientific approaches is their *freedom* - the fact that the way we think of their actions goes beyond the framework of physical causation. My aim at this point will be to approach the concept of a person from a different (but complementary) perspective. I want to ask the following: what kinds of actions make persons the *right* kind of persons who do justice to the “common intentions” of the (Sellarsian) community? By pursuing what kind of *good* can a person reach the “ideal” of personhood? And how does this *ideal* relate to the folk concept of soul that brings us to *folk dualism* more explicitly?

By introducing empirical research in dualist intuitions and folk concept of personal identity I will try to uncover very important nuances in folk conceptualization of persons. The concept of *soul* that I talk about in this context is not meant to be explicitly connected to religious connotations or concrete cultural views, nor do I refer to it as to a concept that is a result of deeper philosophical reflexion. The concept of soul as it is introduced here is simply a result of how *folk* usually understand the term on the everyday basis and the way they use it in the context of talking about certain aspects of human beings. In accord with how I described my opinion about the right method of experimental philosophy, the definition of the concept as I use it here will become clearer with gradual revealing of the results of numerous empirical studies of folk intuitions. The same applies to the concept of *mind* that will be addressed in order to make clear if and to what extent it belongs to the centre of the folk concept of a person and the essential self.

The topic of personal identity that follows the search for the folk concept of soul captures the folk concept of a person in an intriguing way: by asking the

question what is the most essential trait that makes a person who they are deep inside we arrive at the very core of what makes each person the person they *ought to be*. I will point to similarities between this normative dimension of the personal identity problem and the normative dimension revealed in the folk concept of soul and thereby provide a picture of *essential self* in the unity of these two points of view. I will also introduce the results of my own empirical studies in order to illustrate the weight of experimental philosophy and the fact that the results based on studies of Czech population are in accord with the rest of studies I mention here.

3.1 The complexity of folk dualism: Mind vs. soul

Dualism seems to be wired deep into human psychology. As Paul Bloom notes (while referring to Descartes), it is natural for us to think that we are something else (and more) than our bodies.¹⁸² Bloom looks for a natural explanation for this: it should not be surprising that humans have to conceptually differentiate between physical and social realms. It would not be efficient to treat a person in the same way we treat physical bodies. Thus, those individuals who managed to treat other persons in the right way gained evolutionary advantage. Capacities such as mindreading and empathy are adaptive.¹⁸³

Dualistic intuitions are easily observable already in very young children. David Estes refers to research according to which preschool children have “*explicit knowledge of how mental and physical phenomena differ*”. They view mental entities as “*inherently private rather than public*” and as something that cannot be manipulated in the same manner as physical objects.¹⁸⁴ However, dualistic understanding of the world goes even deeper than this obvious differentiation: Bloom

¹⁸² Bloom, P., *Descartes' baby: How the science of child development explains what makes us human*, p. 165.

¹⁸³ Ibid., p. 37.

¹⁸⁴ Estes, D. (2006). Evidence for early dualism and more direct path to afterlife beliefs. *Behavioral and Brain Sciences*, 29(5), 470. Commentary in discussion on: Bering, J. M. (2006). The folk psychology of souls. *Behavioral and Brain Sciences*, 29(5), 453–462; discussion 462–498.

sees children as natural dualists who see the world “*as containing two distinct domains*”, physical and mental, bodies and souls.¹⁸⁵

Dualist thinking shows itself nicely when children are asked about what happens to the dead. A study by Jesse Bering and David Bjorklund¹⁸⁶ revealed that children view a mind of a dead being as still functioning in certain regards: the children watched a puppet performance in which a mouse with certain human traits was devoured by an alligator. Subsequently, the children were asked to ascribe certain biological and psychological states to the killed mouse. They did not ascribe to the dead mouse biological functions such as need to eat or drink. Majority of children even acknowledged that the brain had stopped functioning. However, when asked about mental states such as feeling hunger, thinking and knowing something, children agreed that the dead mouse still has these states of mind.¹⁸⁷ Interestingly, items connected to positive emotions and epistemic states that had something to do with interpersonal relationships (mouse still loves her mom or believes that she is the nicest grownup) were believed to continue after the mouse’s death to the highest degree.¹⁸⁸

As Bloom sums up the results of this research, according to the children “*the soul survives*”.¹⁸⁹ Bloom mentions an anecdote with his son Max who, when he was six years old, described a brain as taking care of hearing, seeing, smelling and most importantly, thinking. However, he expressed a belief that the brain does not feel sad, or love Max’s brother; it helps Max with that, at best, but it is Max *himself* who does these things. Natural view that children have about the brain is that it is a very important organ but still only a tool that helps soul with certain mental operations. Bloom adds that he doubts that adults have a radically different conception of what brain does. They are still surprised when confronted with scientific findings which

¹⁸⁵ Bloom, P., *Descartes’ baby*, p. 169.

¹⁸⁶ Bering, J., & Bjorklund, D. (2004). The natural emergence of reasoning about the afterlife as a developmental regularity. *Developmental Psychology*, 40(2), 217-233.

¹⁸⁷ Bering, J. M., *The folk psychology of souls*, p. 454.

¹⁸⁸ Bering, J., & Bjorklund, D., *The natural emergence of reasoning about the afterlife as a developmental regularity*, p. 226.

Interestingly, the effect is strongest in the case of young children and becomes less prominent with growing age, but nevertheless remains also in adult respondents.

¹⁸⁹ Bloom, P., *Descartes’ baby*, p. 176.

show brain's involvement in thinking about certain topics, especially those connected to the moral realm in certain way.¹⁹⁰ As Bloom concludes, the premise of his book is "*that we are dualists who have two ways of looking at the world: in terms of bodies and in terms of souls. A direct consequence of this dualism is the idea that bodies and souls are separate. And from this follow certain notions that we hold dear, including the concepts of self, identity, and life after death.*"¹⁹¹

Bloom seems to be pointing to something important about dualist intuitions when he mentions *souls*. Indeed, natural dualism does not seem to be simply *mind-body* dualism; Bloom chooses to talk about *body-soul* dualism. However, several studies support the idea that mind and soul are concepts that should be carefully distinguished. Rebekah Richert and Paul Harris explicitly criticise Bloom for not distinguishing properly between the concept of soul and the concept of mind. Such distinction shows itself in many languages, in which speakers distinguish between physical, mental, and spiritual aspects of the self.¹⁹² Richert and Harris support their conceptual theory by referring to the results of two experiments. In the first experiment, children (65 American children aged 4-12 years) were asked about what kind of change takes place after a baby is baptised (whether it is perceptible by senses; whether the change happens on the outside or inside) and which entity changes – the brain, the mind, or the soul of the baby. In general, children tended to answer that the change happens inside the baby, cannot be seen or touched, and that it happens mostly to the soul.¹⁹³

The second experiment focused on how children (45 American participants, 6-12 years old) perceive the difference between the brain, the mind, and the soul. For example, when asked about a change, children tended to say that the brain and the mind change throughout a person's life, but the soul stays the same. What is more, older children (3rd and 5th grade) were more likely to agree that the soul confers identity to a person rather than the brain or the mind.¹⁹⁴ When asked about the

¹⁹⁰ Ibid., p. 170.

¹⁹¹ Ibid., p. 162.

¹⁹² Richert, R. A., & Harris, P. L. (2006). The ghost in my body: Children's developing concept of the soul. *Journal of Cognition and Culture*, 6(3-4), 409-427, pp. 410f.

¹⁹³ Ibid., p. 414.

¹⁹⁴ Ibid., p. 417.

functions of the brain, the mind, and the soul, children tended to ascribe spiritual functions to the soul (telling right from wrong, afterlife, life-giving force, etc.) and cognitive functions to the mind and the brain (feeling, organising thoughts, etc.). Biological functions were mostly ascribed to the brain.¹⁹⁵ In general, the soul was perceived as something that remains unchanged, fulfils spiritual functions, and plays an important role in person's identity - thus as something different from the mind and the brain.

Another experiment by Richert and Harris¹⁹⁶ was performed with adult participants (161 Californian university undergraduate students). Again, the results support the claim that the concept that people have of mind differs from their concept of soul. The mind was generally associated with lifecycle, cognitive and biological functions, whereas the soul was conceptualized in more spiritual terms (afterlife, connection to the higher power, etc.).¹⁹⁷ Even though over one quarter of respondents were not certain about the soul's existence, most of them claimed that it survives after death. The soul was perceived as a more constant entity than the mind.¹⁹⁸

Richert and Harris thus arrived at a conclusion that most respondents do not subscribe to a simple dualist view. One of the theories that Richert and Harris offer as an explanation for participants' concept of soul is a kind of essentialism: "*...it is plausible that people think of human beings as having an invariant essence that confers a stable identity despite the various transformations brought about by the processes of growth and ageing. To the extent that this essence is preserved despite those transformations, it may be feasible to imagine its existence independent of that biological cycle. On this view, the human lifecycle is no more than the temporary embodiment of something more permanent.*"¹⁹⁹

On the other hand, respondents could have been influenced by their cultural environment (English-speaking, Christian communities) where the "soul discourse"

¹⁹⁵ Ibid., pp. 420f.

¹⁹⁶ Richert, R. A., & Harris, P. L. (2008). Dualism revisited: Body vs. mind vs. soul. *Journal of Cognition and Culture*, 8(1-2), 99-115.

¹⁹⁷ Ibid., pp. 106ff.

¹⁹⁸ Ibid., pp. 104ff.

¹⁹⁹ Ibid., p. 112.

is very common.²⁰⁰ However, we can find similar distinction between the body, the mind, and the soul also in other, distinct cultures. Richert and Harris mention two societies - the Vezo of Western Madagascar and the Lohorung of Eastern Nepal, who use different terms in order to speak about three different aspects of a person. In the case of the Vezo society, the “*term vata corresponds quite closely to the English term ‘body’*. The terms *say* and *fanahy*, in contrast, refer to noncorporeal aspects of the person.” The term *say* is used to refer to “*capacity for intelligent and competent behavior*” and thus corresponds to English term “mind”. The term *fanahy* refers to “*a person’s social disposition*” – characteristics like generosity, anger, friendliness and animosity. A person can have good *fanahy* independently of *say*, and *fanahy* is also believed to be able to leave the body (esp. during sleep and after death). The other society, the Lohorung Rai of East Nepal, use terms *niwa* and *lawa*. *Niwa* corresponds to the mind, since it is used to talk about reasoning and cognitive capacities, whereas *lawa* refers to an entity which comes to a child already in the womb and which is capable of leaving the body (again esp. during dreaming and after death).²⁰¹

These similarities point to a possibility that the concept of soul, which is essential for person’s identity, is widespread across societies or even universal.²⁰² This leads certain authors to the idea that such a concept could be further studied by using evolutionary approaches. Jesse Bering suggests that “*the standard architecture of ancestral human minds was co-opted by natural selection to create the functional*

²⁰⁰ Ibid., pp.112f.

²⁰¹ Ibid., p. 114.

²⁰² At least in societies nowadays. I am aware that the concept of soul has been developing throughout human history from concepts very distant from how we understand soul today (my thanks go to Vojtěch Hladký for motivating me to delve little deeper into this topic). See e.g. Furley, D. J. (1956). The early history of the concept of soul. *Bulletin of the Institute of Classical Studies*, 3, 1-18; or Chlup, R. (Ed.). (2007). *Pojetí duše v náboženských tradicích světa* [The conception of soul in the religious traditions of the world]. Prague: DharmaGaia.

The complete picture (especially for purposes of evolutionary speculations) would require historically motivated thematization of the concept of soul. Due to the main focus of my thesis (x-phi focusing on societies living nowadays) I do not include such analyses in my thesis. I believe, however, that experimental philosophers should consider using these historical approaches, depending on how general claims they wish to make based on results of their experiments.

illusion of an intelligently designed, immortal soul that was under nearly unbreakable moralistic contract with the natural world."²⁰³

More recent study on 206 American university students by Stephanie Anglin²⁰⁴ gives further support to the proposed structure of folk dualism. The participants were asked about the location of the self, the mind, and the soul in the body and about their own definitions of these entities. Most respondents located the soul in the heart or chest,²⁰⁵ defined it as *essence, inner being, or immaterial self*, and connected the soul also to conscience, morals, and emotions.²⁰⁶ The mind, on the other hand, was located in the brain or head²⁰⁷ and defined as thoughts and consciousness.²⁰⁸

All these studies explicitly addressing the difference between the concept of mind and soul suggest that these entities play quite different roles in folk dualism. The soul seems to be the entity that is more essential to persons and more connected to the moral domain, while the mind is connected to morally more neutral cognitive and intellectual abilities. This might become clearer as I address studies that look at this difference less explicitly: via the problem of personal identity I will show how particular categories of traits are viewed within this context.

3.2 Personal identity and the essential moral self

Another battery of studies that reveal further aspects of this natural dualism present in everyday human thinking comes from the field of experimental philosophy.²⁰⁹

²⁰³ Richert, R. A., & Harris, P. L. (2008). *Dualism revisited: Body vs. mind vs. soul*, p. 115. Quote from: Bering, J. M. (2006). *The folk psychology of souls*, p. 461.

²⁰⁴ Anglin, S. M. (2014). I think, therefore I am? Examining conceptions of the self, soul, and mind. *Consciousness and Cognition*, 29, 105-116.

²⁰⁵ Ibid., p. 109.

²⁰⁶ Ibid., p. 110.

²⁰⁷ Ibid., pp. 109f.

²⁰⁸ Ibid., pp. 110f.

²⁰⁹ It is getting harder and harder to isolate psychological studies from studies made by researchers who are philosophers by training. Experimental philosophy is multidisciplinary and attracts many psychologists and social scientists who, in turn, inspire experimental philosophers.

Nina Strohminger and Shaun Nichols²¹⁰ conducted five experiments on personal identity (79 to 318 respondents, adult Americans), and the results support the hypothesis of “the essential moral self”- intuition of most participants seems to be that *moral traits*²¹¹ constitute the core of personal identity - the “true self” of a person. Other kinds of traits, esp. physical traits, can change without seriously distorting the person’s identity.

In the first experiment, the respondents were asked to consider a scenario in which a person called Jim undergoes partial brain transplant. Then various possibilities of how Jim changes were introduced (visual object agnosia, autobiographical amnesia, apathy, and loss of moral conscience). Participants were supposed to determine to what extent these changes impacted Jim’s identity. Cognitive deficit wasn’t considered to lead to identity change, loss of autobiographical memory had a larger impact, and loss of moral faculty lead to the most vivid perceived change in personal identity.²¹²

²¹⁰ Strohminger, N. & Nichols, S. (2014). The essential moral self. *Cognition*, 131(1), 159-171.

²¹¹ It is important to make clear what is meant by the word “moral” within the context of these studies, since this is often not clear. In some cases, the category of traits labelled as “moral” is determined by the way coders blind to the aims of the study categorize different concrete traits (as in some of the studies by Strohminger and Nichols), in other studies the questions are more abstract (asking about moral values without further specifications). However, in general, the traits that are labelled as “moral” for short stand for the traits that play a role in interpersonal behaviour (Jirout Košová, M., *Skúmanie významu experimentálnej filozofie skrze koncept osobnej identity*, pp. 596f). The connection of morality to interpersonal relationships is therefore crucial here. This should not be surprising, after all, given the current theories promoting the idea that “*the function of morality is to promote cooperation*”. See Curry, O. S., Alfano, M., Brandt, M. J., & Pelican, C. (2020, June 9). Moral molecules: Morality as a combinatorial system. Retrieved from <https://doi.org/10.31219/osf.io/xnstk> (preprint version 4), pp. 3ff.

One should still bear in mind that in each study specific vignettes are used, and only together they can provide an idea what “moral” means in the wider context of x-phi studies. Again, this is a situation typical for x-phi and one still has to remember that the folk are theory-lite, and we get to their concepts gradually, by forming a larger picture from a larger number of different studies.

²¹² Strohminger, N., & Nichols, S., *The essential moral self*, p. 161.

The scenario in the second experiment introduced the idea of a pill capable of permanently changing one particular part of the person's mind. The respondents evaluated how much a person would change after having one of 62 cognitive and behavioural traits (perception, desires and preferences, memories, and morality) permanently altered. Again, change in moral traits was considered to lead to the most dramatic perceived change in personal identity. When it came to memories, they didn't seem to be particularly important for personal identity as such but rather depending on what they are memories of – memories of some practical knowledge were rated low, whereas episodic memories connected to the social and personal context (memories of traumatic events, precious memories of time spent with family members) were rated as important. As the authors of the studies conclude, “*memory is not important unto itself, but rather for the connections it affords us to our socio-moral core.*”²¹³

In the third study the authors aimed at the relationship between the essential moral self hypothesis and the concept of soul. They asked participants about what characteristics would move with a person's soul if it could move from one body to another. Moral traits (e.g. being honest, evil, conscientious, a coward) were most significantly associated with the soul (but again, so were traumatic and personal memories). The tendency of the respondents to place certain traits in the new body didn't differ depending on their religiosity or belief in souls.²¹⁴

In the fourth study that was concerned with the idea of reincarnation the authors focused on moral character and personality. Respondents were supposed to imagine that reincarnation really happens and that a person's “true self” (real identity) is preserved throughout the whole process. They were further asked to look at pairs of characteristics (one item strongly related to morality, the other not, e.g. honest vs. smart) and choose which one of them was more likely to have been preserved during a person's reincarnation. All moral items were more likely to be chosen over personality items. Belief in reincarnation, soul, or religiosity didn't show as an important factor in preferring moral traits.²¹⁵

²¹³ Ibid., p. 162f.

²¹⁴ Ibid., p. 164f.

²¹⁵ Ibid., p. 165f.

In the last study, respondents were supposed to imagine that they had met an old friend after forty years. Then they were presented with a list of changed traits and asked to rate to what extent each change impacted their friend's identity (his true self). Again, change of moral characteristics was assessed as having the most significant impact. Personality traits ended in the second place while after came basic cognition, memories, preferences, and perception.²¹⁶

Strohminger and Nichols conclude that lay theories of personal identity are based on moral considerations – more precisely, what is important about a person is not so much some distinctive trait (which would help us “*pick the person from the crowd*”). Rather, the essential ingredient of a person seems to be the set of characteristics closely connected to moral conduct and interpersonal relationships. These are characteristics which are usually very general, like empathy and possession of moral compass. This is not surprising, since “*moral traits are a reliable predictor for how individuals will fare as potential partners for cooperation and affiliation.*”²¹⁷

The authors came to similar conclusions also in their study based on real-life scenarios – they conducted an online study with the relatives of people suffering from neurodegenerative diseases. The respondents were asked to judge the impact of different symptoms on relationship with their affected close ones and also answer questions about personal identity. Change in moral traits and interpersonal dispositions proved to have significantly greater impact on the continuity of personal identity than other kinds of changes brought about by neurodegeneration.²¹⁸

Jesse Prinz and Shaun Nichols²¹⁹ revealed the importance of moral continuity for preserving personal identity in their studies launched in 2009. Memory, agency, or narrative ability showed as significantly less important for the folk concept of diachronic identity than morality. They also explain their results by pointing to the

²¹⁶ Ibid., p. 166ff.

²¹⁷ Ibid., pp. 168f.

²¹⁸ Strohminger, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science*, 26(9), 1469 - 1479.

²¹⁹ Prinz, J., & Nichols, S. (2016). Diachronic identity and the moral self. In J. Kiverstein (Ed.), *The Routledge handbook of philosophy of the social mind* (pp. 449-463). Abingdon: Routledge.

fact that morality is a social phenomenon. Moral values of the community are the measure of our success as social beings.²²⁰ Personal identity is defined in the context of how people view it in everyday social interactions and practices.²²¹

What is more, further studies showed that the negative moral change is the crucial kind of change that disrupts personal identity in the eyes of most respondents. When a good person turns into morally bad person, his personal identity is perceived by the respondents as broken more radically than when the opposite is the case.²²² Strohminger, Knobe, and Newman²²³ suggest that apart from a broader concept of self, the folk also have a narrower concept of *true self*. Concept of self is much wider and variable, while the concept of *true self* has specific emphasis on positive moral features and is even stable across cultures.²²⁴ Different cultures from around the world express the same view, namely that the *true self is good*. This cross-cultural similarity shows itself on a “*more abstract level*”, since being good may mean different things in different cultures. The idea that “*the true self is calling us to morally good actions*” seems to be genuinely universal, though.²²⁵ Authors suggest that this has to do with our nature as social beings and perhaps our natural essentialism.²²⁶

Heiphetz, Strohminger, and Young²²⁷ brought attention to the role of social bonds by showing that they seem to be crucial for the folk concept of personal identity. Participants in their studies judged belonging to a group as important for them and statistical analyses revealed the relationship between the perceived importance of community and the type of moral belief they judged to be more important for preserving personal identity. Changes to widely shared moral beliefs

²²⁰ Ibid., p. 463.

²²¹ Ibid. p. 449.

²²² Tobia, K.P. (2015). Personal identity and the Phineas Gage effect. *Analysis*, 75(3), 396-405.

²²³ Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*, 12(4), 551-560.

²²⁴ Ibid., pp. 552ff.

²²⁵ Ibid., p. 554.

²²⁶ Ibid., pp. 556f.

²²⁷ Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The role of moral beliefs, memories, and preferences in representations of reality. *Cognitive Science*, 41(3), 744-767.

elicited more radical identity change than changes in controversial moral beliefs, because widely shared beliefs are more closely connected to relationships in the community.²²⁸

The importance of interpersonal traits in judgements about personal identity comes forward also in the research conducted by Kevin Tobia²²⁹ who focused on the direction of hypothetical change. In his study he broadened the classical “Phineas Gage” scenario and found that respondents presented with the reversed version of the scenario (moral improvement) tended to claim that he is still the same person after the accident, whereas in the original scenario (moral deterioration) the opposite tendency was the case. Both changes were equally radical, but the negative version lead more often to conclusions that the identity of the person in question was broken. Tobia suggests that the reason why responses to opposing scenarios differ so significantly is due to the fact that the self is viewed as essentially good.²³⁰ In his earlier paper²³¹ Tobia referred to other examples (Parfit’s *nobleman* thought-experiment and different examples from pop-culture and literature). His conclusion is that *normativity* connected to the social dimension is to be viewed as a crucial aspect of the folk concept of personal identity.²³²

Adult respondents in the study of Heiphetz, Strohminger, Gelman, and Young²³³ judged the change from good moral beliefs to bad moral beliefs to have a significantly more serious impact on preserving personal identity than the change in the opposite direction. The difference was found to be connected to the perceived influence of both types of changes on friendships.²³⁴ Also child respondents reported, similarly as adults, that the person would change more after a change in their widely

²²⁸ Ibid., p. 758.

²²⁹ Tobia, K.P. (2016). Personal identity, direction of change, and neuroethics. *Neuroethics*, 9(1), 37-43.

²³⁰ Ibid., pp. 39f.

²³¹ Tobia, K.P., *Personal identity and the Phineas Gage effect*.

²³² Ibid., p. 404.

²³³ Heiphetz, L., Strohminger, N., Gelman, S., & Young, L. (2018). Who am I? The role of moral beliefs in children’s and adults’ understanding of identity. *Journal of Experimental Social Psychology*, 78, 210-219.

²³⁴ Ibid., p. 216.

shared moral beliefs took place, in comparison to change in controversial moral beliefs, memories, or preferences.²³⁵

Apart from moral traits, memory is another crucial pillar of personal identity and has always been considered in this context (the classic example is John Locke)²³⁶ and was discussed also in findings of contemporary authors.²³⁷ Despite the role of memory, studies exploring the difference between the importance of moral traits and other categories of traits for preserving personal identity show that moral traits surpass memory in the context of personal identity²³⁸, although not consistently in all studies.²³⁹

I have already mentioned that some authors tend to explain these and other similar results – that people seem to consistently believe in the good true selves - by suggesting that people are naturally essentialists. *Psychological essentialism* is the tendency to view entities as having an invisible essence hidden behind visible superficial features.²⁴⁰ De Freitas et al. propound that the belief in self bears crucial characteristics of psychological essentialism. Characteristics that are usually ascribed to the true self are viewed as “*immutable, discrete and inherent*”.²⁴¹ We have seen that the moral traits that are so closely connected to the true self are the central pillar of personal identity, thus it is apparent that they are indeed viewed as immutable and essential aspect of a person.

²³⁵ Ibid., pp. 211-214.

²³⁶ Locke, J. (2009). *Essay Concerning Human Understanding*. WLC books. (Originally published in 1690)

²³⁷ E.g. Nichols, S., & Bruno, M. (2010). Intuitions about personal identity: An empirical study. *Philosophical Psychology*, 23(3), 293-312.

²³⁸ I have already mentioned Prinz, J., & Nichols, S., *Diachronic identity and the moral self*; Strohming, N., & Nichols, S., *The essential moral self*; and Strohming, N., & Nichols, S., *Neurodegeneration and identity*.

²³⁹ As in Heiphetz, L., Strohming, N., & Young, L. L., *The role of moral beliefs, memories, and preferences in representations of reality*.

²⁴⁰ De Freitas, J., Sarkissian, H., Newman, G. E., Grossman, I., De Brigard, F., Luco, A., Knobe, J. (2018). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science*, 42(S1), 134-160, p. 138.

²⁴¹ Ibid.

The true self studies mentioned above meet nicely with the studies that aim at essentialism since they all show that the inner essence is viewed as inherently good. De Freitas et al. managed to gain substantial evidence for this tendency by aiming at misanthropes²⁴² and three different interdependent cultures (comparing U.S. sample to Russia, Colombia, and Singapore).²⁴³ In the case of misanthropy, the results suggested that while people who showed tendency towards misanthropy were more likely to be pessimistic in their predictions of future actions of an agent, they nevertheless exhibited the true self bias – they still believed that the essence of the person in question was good.²⁴⁴ Belief in a good true self showed to be consistent even across different cultures. Participants from Russia, Colombia, and Singapore, similarly to the American respondents, tended to attribute positive moral changes to the true self in a higher degree than when the changes were negative. Magnitude of the good true self effect didn't differ significantly across countries.²⁴⁵ Further studies suggest that even the members of the outgroups that are usually viewed with animosity are believed to possess good true selves.²⁴⁶

Psychological essentialism seems to be a wider tendency to ascribe positive inner essence not only to the human self but also to other, non-human entities.²⁴⁷ De Freitas et al. refer to studies that show that people view identity of entities such as institutions, groups, or texts in the light of their belief that these entities have normatively good essences.²⁴⁸ Apart from normativity, there also might be a kind of teleology in play here,²⁴⁹ as new experimental results show that people view essences

²⁴² Ibid., pp. 140-145.

²⁴³ Ibid., pp. 146-151.

²⁴⁴ Ibid., pp. 141, 144.

²⁴⁵ Ibid., p. 151.

²⁴⁶ De Freitas, J., & Cikara, M. (2018). Deep down my enemy is good: Thinking about the true self reduces intergroup bias. *Journal of Experimental Social Psychology*, 74, 307-316.

²⁴⁷ De Freitas, J., Tobia, K., Newman, J. E., & Knobe, J. (2017). Normative judgements and individual essence. *Cognitive Science*, 41(S3), 382-402, p. 385.

²⁴⁸ Ibid., p. 400.

²⁴⁹ Jirout Košová, M., Kopecký, R., Oulovský, P., Nekvinda, M., & Flegr, J. (in press). My friend's true self: Children's concept of personal identity. *Philosophical Psychology*. Advance online publication (preprint version 3) retrieved from psyarxiv.com/uwa59, p. 17.

of various entities via their purposes.²⁵⁰ People simply consider a specific purpose of the entity in question, i.e. the purpose of the music band is to make good-quality music, purpose of scientific papers is to carry valuable scientific information, and the purpose of a person is to pursue moral goodness.²⁵¹

What is more, the theory of psychological essentialism is also in accord with findings concerning child behaviour.²⁵² Infants seem to prioritize internal features of an object when considering its behaviour, and this happens precisely in cases when the objects exhibit self-generated motion.²⁵³ Persons are a clear example of entities whose behaviour is self-generated, thus we can expect that essentialism is the right framework to apply to human agents.

Psychological essentialism thus seems to be a plausible explanation of the experimental results concerning the true self bias. Another question is to what extent people mean it when they say “it is not the same person anymore” in these cases of moral change. De Freitas and colleagues maintain the view that people truly believe that with moral deterioration comes the end of the person in question in the sense of numerical identity.²⁵⁴ On the other hand, Starmans and Bloom argue that people only refer to a significant change, not the annihilation of the person.²⁵⁵ However, there are studies such as that of Tobia mentioned earlier,²⁵⁶ that successfully distinguish between numerical identity and similarity²⁵⁷ and still reveal the true self effect.²⁵⁸

²⁵⁰ Rose, D., & Nichols, S. (2019). Teleological essentialism. *Cognitive Science*, 43(4), e12725. doi:10.1111/cogs.12725

²⁵¹ De Freitas, J., Tobia, K., Newman, J. E., & Knobe, J., *Normative judgements and individual essence*, p. 397.

²⁵² See e.g. Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. New York: Oxford University Press.

²⁵³ Newman, G. E., Herrmann, P., Wynn, K., & Keil, F.C. (2008). Biases towards internal features in infants' reasoning about objects. *Cognition*, 107(2), 420-432.

²⁵⁴ De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2018). Moral goodness is the essence of personal identity. *Trends in Cognitive Sciences*, 22(9), 739-740.

²⁵⁵ Starmans, C., & Bloom, P. (2018). Nothing personal: what psychologists get wrong about identity. *Trends in Cognitive Sciences*, 22(7), 566-568.

²⁵⁶ Tobia, K. P., *Personal identity and the Phineas Gage effect*.

²⁵⁷ Dranseika, V. (2017). On the ambiguity of 'the same person'. *AJOB Neuroscience*, 8(3), 184-186.

Anyway, I believe that certain vagueness is inseparable from the folk intuitions, as we have already seen in the case of the free will concept. These debates lose their appeal when we realise that folk don't attempt to come up with elaborated theories – the folk are “theory-lite”. They simply express intuitions that stem from the fact that humans are essentially social beings and use concepts that make the social world of interpersonal relationships easier to handle. The fact that the concept of true self seems to have a normative and teleological character only strengthens this point: it is useful for us to view others as essentially good deep inside (in other words, capable of fulfilling the real purpose of human being), because this makes it easier to form and maintain positive social relationships. What is more, claiming that someone is “not the same person anymore” when they turn bad might serve as a kind of social sanction by means of taking the person's identity away from them.²⁵⁹

Based on the research listed above and its conceptual analysis, it is possible to claim that the *essential moral self* and the *true self* concepts are very closely related to the *soul* concept.²⁶⁰ Respondents in both the experimental philosophy and psychological essentialism studies associated the *true self* of a person with traits that have to do with moral conduct and interpersonal capabilities rather than cognitive abilities connected to the mind or physical aspects of the person. What is more, when asked about souls moving from one body to another in the Strohminger and Nichols study,²⁶¹ the traits that were rated as most likely to move with the soul were precisely of the same type as characteristics associated with the essential moral self and identity of a person in other experiments. On the top of that, Richert and Harris showed in their experiments that apart from being connected to moral dimension, the soul is also viewed as a constant and unchanging entity connected to a person's

²⁵⁸ Jirout Košová, M., Kopecký, R., Oulovský, P., Nekvinda, M., & Flegr, J., *My friend's true self: Children's concept of personal identity*, p. 8.

²⁵⁹ Ibid., p. 17.

²⁶⁰ Note that this is not the case for the *wider self* concept: the word “self” as we use it in everyday language does not overlap with the concept of soul but exhibits certain ambiguity and approaches definitions associated rather with the mind. See Anglin, S. M., *I think, therefore I am? Examining conceptions of the self, soul, and mind*, p.114.

²⁶¹ Strohminger, N., & Nichols, S., *The essential moral self*, pp. 164f.

identity.²⁶² The essential core of a person, the “entity” the person really is, seems to be the soul carrying the moral traits which are important in social context of interpersonal relationships.

In order to better demonstrate the appeal of experimental results for this conclusion, I will refer to my own experimental research and thus show in greater detail how folk intuitions are uncovered.

3.3 The essential moral self and folk dualism: Experimental studies

The following two studies are concrete and more detailed examples of the findings described in the previous section. Thanks to the fact that I have had an opportunity to take part in original research²⁶³ I was able to see directly that the concept of “essential moral self” indeed seems to be present in people universally and that it bears undeniable connection to the concept of soul. I will describe both studies and point to the connection between them that reveals the crucial aspect of folk dualism as I explore it in the context of this thesis.

In order to contribute to the topic of personal identity in experimental philosophy we decided to conduct an interview study with children.²⁶⁴ These studies are still quite rare, especially in the non-English speaking countries.

In June 2017 we interviewed 217 Czech children and teenagers (56.4% female, age range 6-15 and average age of 11 years) as a part of public event popularising science. We interviewed each child separately and introduced them to a scenario in which their friend, someone they know, or some person in general (different groups of respondents for each of these three versions) undergoes various changes after entering a special sci-fi chamber. Changes from six categories were introduced, and both negative and positive changes were included for each item of

²⁶² Richert, R. A., & Harris, P. L., *The ghost in my body: Children's developing concept of the soul*, p. 417; and Richert, R. A. & Harris, P. L., *Dualism revisited: Body vs. mind vs. soul*, pp. 104ff.

²⁶³ Supported by the Charles University Grant Agency, project GA UK 925416 (“Concept of personal identity from the perspective of experimental philosophy”) in which I was the principal researcher.

²⁶⁴ Jirout Košová, M., Kopecký, R., Oulovský, P., Nekvinda, M., & Flegr, J., *My friend's true self: Children's concept of personal identity*.

each category: morality (the person becomes crueller or nicer, stops loving their close ones or accepts as friend someone they didn't like before), physical appearance (the person becomes more ugly or more beautiful), cognition (the person becomes more stupid or smarter), character (the person becomes more lazy or more industrious), episodic memory (the person forgets all his life memories or gains a supermemory), and perception (the person becomes blind or gains much better eyesight). The children were asked to judge how much each of the changes would affect the person's identity core ("*the most crucial aspect of the person which makes them who they really are deep inside*") on a 7-point scale (0 – "*they are still the same person and their most crucial aspect remains intact*"; 6 – "*they are not the same person anymore and have lost their most crucial aspect*"). The scale was indicated by circles growing in size.²⁶⁵

The results of t-tests are in accord with the studies introduced in the previous section. Change in moral traits was rated by the respondents as causing significantly larger overall change in personal identity than change in any of the other categories of traits. Change of morality with the largest perceived impact on personal identity was followed by memory, cognition, character, and perception. Physical traits ended up in the last place with the lowest score (see fig. 1).²⁶⁶

²⁶⁵ Ibid., pp. 8ff.

²⁶⁶ Ibid., pp. 12f. The graphs used here show means of the original, untransformed values (see appendix of the cited preprint).

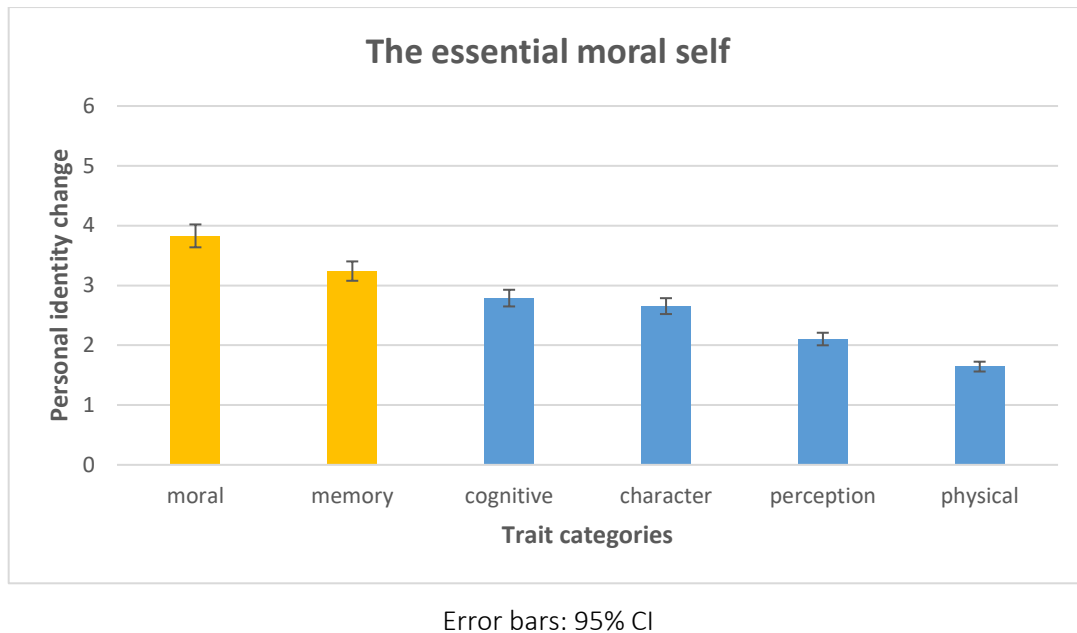
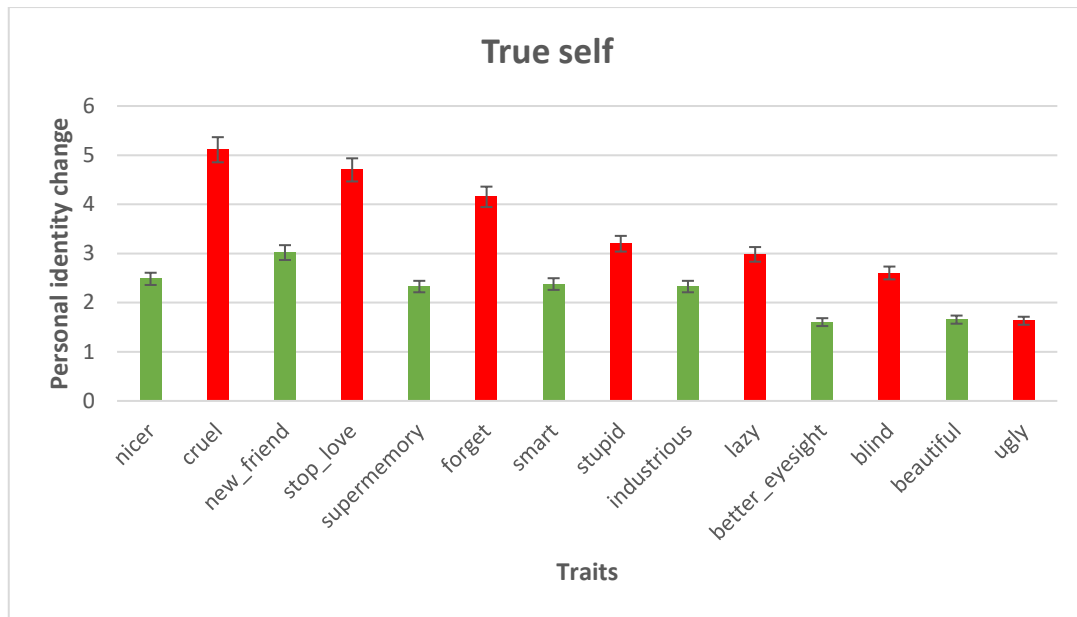


Fig. 1

Mean values of the original scores of the perceived change of personal identity for all participants (y-axis) as they differ across six categories of change (x-axis). Change in moral traits was rated by the respondents as causing significantly larger overall change in personal identity than change in any of the other categories of traits. I indicate here the categories usually connected to the “essential moral self” concept in yellow colour, while the rest of the categories are marked in blue.

Changes in negative direction proved to be viewed as having significantly greater impact on personal identity change than changes in positive direction in almost all categories, except physical. The differences between positive and negative versions were most salient in the case of moral category. Memory followed with smaller but still highly significant difference, while in the case of physical appearance both versions received almost equal rating (see fig. 2).²⁶⁷

²⁶⁷ Ibid., p. 13.



Error bars: 95% CI

Fig. 2

Mean values of the original scores of the perceived change of personal identity for all participants (y-axis) as they differ across fourteen items of change (x-axis). The negative changes are marked in red colour, while the positive changes are marked in green. Changes in negative direction proved to be viewed as having significantly larger impact on personal identity change than changes in positive direction in almost all categories, except physical.

In one of the further exploratory analyses we were also aiming at the effect of age. We tested the differences between scores of moral and non-moral categories across three age groups and the differences proved to be significant between the age groups of 6-8 years and 9-11 years, and the age groups of 6-8 years and 12-15 years.²⁶⁸

We performed the same analyses in order to explore the effect of scenario and found significant differences between the relative importance of moral traits in comparison to non-moral traits: participants who were asked about the change of their *friend* rated moral traits to be significantly more important than non-moral traits in comparison to participants who were asked about the change of *some person in general*.²⁶⁹

²⁶⁸ Ibid., p. 14.

²⁶⁹ Ibid., p. 15.

The results suggest that the concept of the *essential moral self* and *true self* are already present in children and that the importance of the moral category in preservation of personal identity seems to grow with age. What is more, the more personal the scenario, the more important the moral traits seem to be in comparison to the other categories of traits.²⁷⁰

In order to determine the connection between the folk concept of personal identity and the folk concept of soul we also performed an online dualism study²⁷¹ with Czech adult respondents. The study was a part of a larger study that was composed of several questionnaires on different topics. To make sure that the respondents paid attention to the questions we introduced two attention checks (one before and one after the dualism questionnaire). Only the answers from the respondents who answered both attention checks correctly were subject to final analysis.

In the end, we managed to gain 2996 respondents (average age 34 years, 66% females, 56% reported to be non-believers). We aimed at the difference between the concepts of brain, mind, and soul, so we asked our respondents to consider which human traits and abilities are dependent on the brain, on the mind, and on the soul. We asked whether the *brain*, *mind* and *soul* (within-subject model, randomized order) are important for the following abilities (randomized order) on a 5-point scale (from *definitely not* to *definitely yes*): for *an ability to remain the same person throughout one's life*; for *one's eyesight*; for *an ability to love one's close ones*; for *an ability to do mathematical computations*; for *an ability to remember one's credit card pin number*; for *an ability to remember one's close ones*; for *an ability to feel compassion*; for *an ability to distinguish between good and evil*; for *an ability to*

²⁷⁰ Ibid., pp. 15ff.

²⁷¹ Michaela Jirout Košová, Robin Kopecký, Pavel Oulovský, and Jaroslav Flegr. Robin Kopecký, Pavel Oulovský, and Jaroslav Flegr helped me with formulating the questions for this study by providing very useful suggestions and ideas. The questions were inspired by the previously mentioned work on dualism by Rebekah Richert and Paul Harris. The study is not yet published. Here I provide preliminary results of the study that should, nevertheless, illustrate my point. I provide more detailed information on statistics than in the case of previous study since no preprint with further details is available for the dualism study at the time. Statistical analyses reported here were performed by me.

move one's body; for an ability to realise where one is; for an ability to desire something. In the case of the mind and the soul questions we added also a possibility to express disbelief in these entities in order to avoid forcing the respondents into unnatural answers.²⁷²

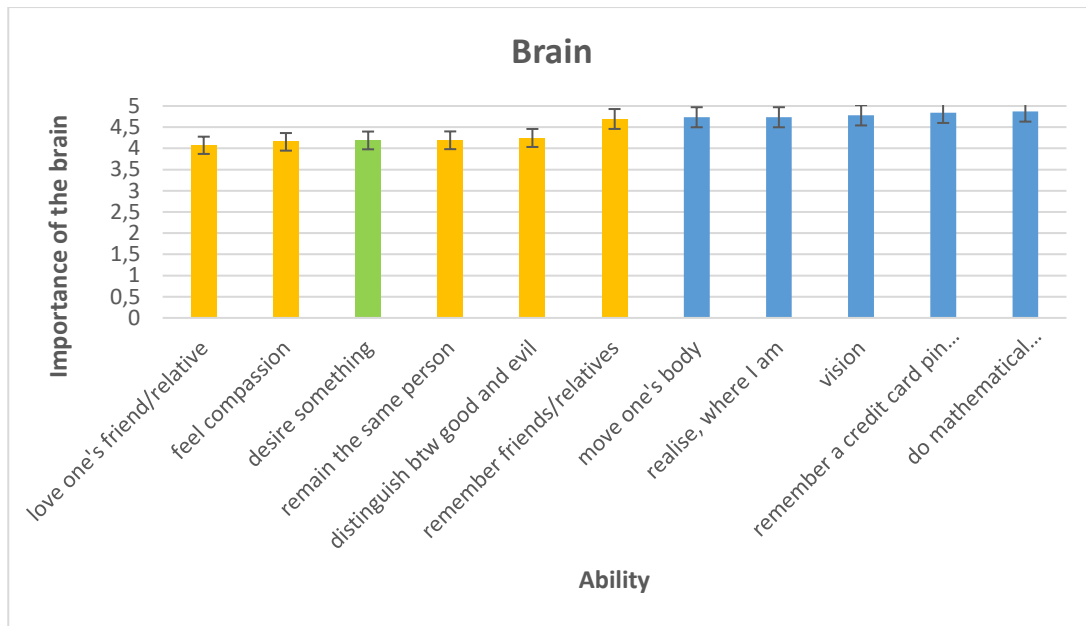
We hypothesised that the subjects will be significantly less willing to ascribe importance to the brain in the case of the “essential moral self” traits and abilities than in the case of purely intellectual and physical abilities. We expected the opposite pattern in the answers to the soul question. We also expected that the answers to the mind question will be more ambiguous and structurally different from the answers to the soul question.

The results²⁷³ proved that our hypotheses were pointing in the right direction. Even though the respondents clearly acknowledged the brain as a basis for all traits and abilities, in the case of items connected to the “essential moral self” they rated the importance of the brain as significantly lower²⁷⁴ than in the case of the intellectual/physical (non-moral) items (see fig. 3).

²⁷² In the case of mind 135 respondents (less than 5%) and in the case of soul 729 (around 25%) chose the disbelief option (in average for each trait/ability). Some of the respondents did not choose the disbelief option consistently across all traits and decided to ascribe some traits to the entities in question.

²⁷³ I performed one-tailed paired t-tests between means for moral (essential moral self) and non-moral categories (leaving out the neutral “desire” item) for the brain and soul questions, and two-tailed paired t-tests for the mind question.

²⁷⁴ $M=-0.521$, $SD=0.736$, $t(2943)=-38.44$, $p<0.0001$.



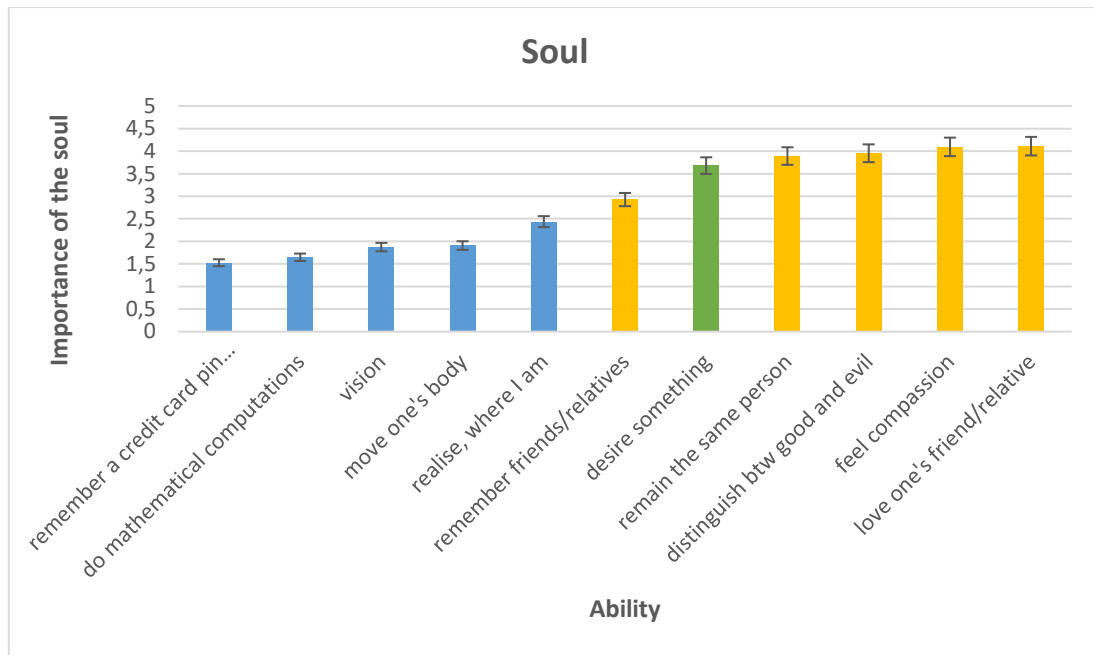
Error bars: 95% CI

Fig. 3

Mean values of the original scores of the perceived importance of the brain (y-axis) as they differ across eleven items (x-axis). The traits usually associated with the “essential moral self” are marked in yellow colour, while the intellectual/physical traits are marked in blue. The neutral psychological trait (*desire something*) is marked in green. The traits are ordered from the lowest to the highest average score. In the case of items connected to the “essential moral self” respondents rated the importance of the brain as significantly lower than in the case of the intellectual/physical items.

In the question about the importance of the soul the pattern was exactly the opposite. Despite respondents ascribed lower importance to the soul overall, the perceived importance of the soul for the “essential moral self” traits was significantly higher²⁷⁵ than for the intellectual/physical traits (see fig. 4).

²⁷⁵ M=1.91, SD=0.96, t(2057)=90.25, p<0.0001 (p-value approaching 0).



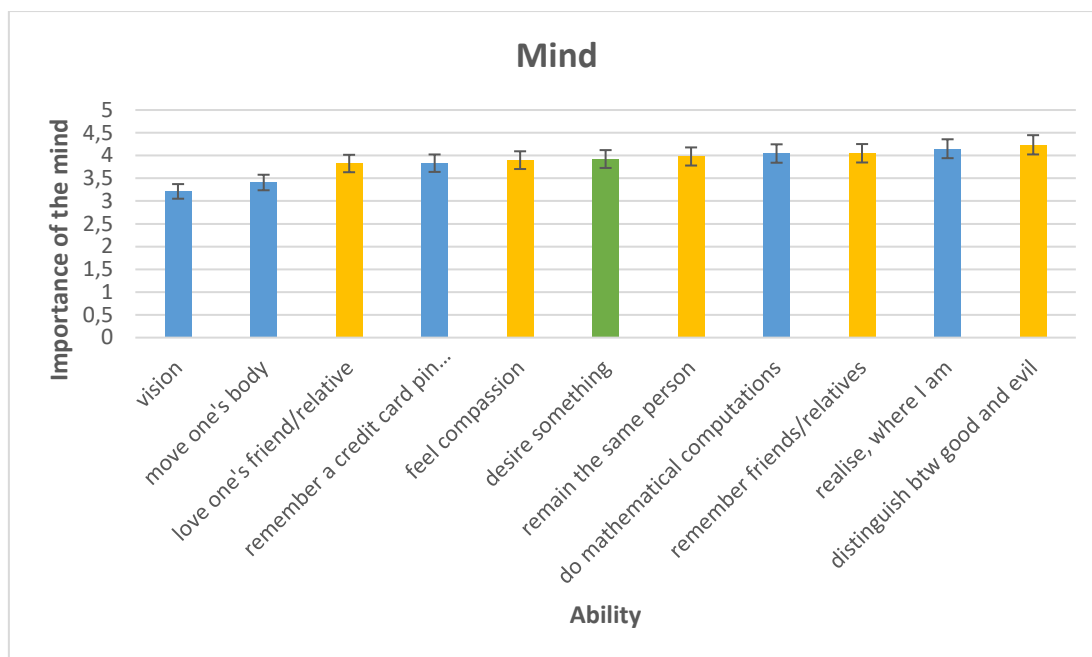
Error bars: 95% CI

Fig. 4

Mean values of the original scores of the perceived importance of the soul (y-axis) as they differ across eleven items (x-axis). The traits usually associated with the “essential moral self” are marked in yellow colour, while the intellectual/physical traits are marked in blue. The neutral psychological trait (*desire something*) is marked in green. The traits are ordered from the lowest to the highest average score. In the case of items connected to the “essential moral self” respondents rated the importance of the soul as significantly higher than in the case of the intellectual/physical items.

The pattern of the answers in the case of the mind proved to be much more ambiguous. Even though there was a significant difference between the “essential moral self” traits and the intellectual/physical traits, it was much lower than in the case of the brain and soul (see fig. 5).²⁷⁶

²⁷⁶ $M=0.26$, $SD=1.02$, $t(2711)=13.29$, $p<0.0001$.



Error bars: 95% CI

Fig. 5

Mean values of the original scores of the perceived importance of the mind (y-axis) as they differ across eleven items (x-axis). The traits usually associated with the “essential moral self” are marked in yellow colour, while the intellectual/physical traits are marked in blue. The neutral psychological trait (*desire something*) is marked in green. The traits are ordered from the lowest to the highest average score. The pattern of the answers in the case of the mind proved to be ambiguous, with the “essential moral self” traits being mixed with the intellectual/physical traits.

In order to further test the difference between the concepts of mind and soul we asked the participants to ascribe certain traits and abilities (“*In case you believe in mind/soul, would you describe it in the following way?*”) to either mind or soul (between-subject design) on a 5-point scale (from *definitely not* to *definitely yes*), with extra disbelief and “don’t know” answer options.²⁷⁷ The items read as follows (randomized order): The mind/soul is *independent from the body*; *immortal*; *after physical death it gets separated from the body and continues to exist*; *it can reincarnate (move) to a new body*; *exists out of space*; *is immaterial*; *multiple*

²⁷⁷ There were over 1400 participants in each of the conditions. Around 11% chose the “don’t know” option in the soul condition and around 10% in the mind condition (in average for a trait). Around 22% chose the disbelief option in the soul and around 5% in the mind condition.

(souls/minds) can exist in one body; also some animals have (mind/soul); in the future it will be possible to upload it to a computer.

As we hypothesised, there turned out to be visible differences in how respondents ascribed different traits and abilities to the mind and soul.²⁷⁸ Continued existence, immortality, ability to reincarnate, independence from space, independence from the body, and immateriality (in the order of significance) were ascribed to the soul to a significantly higher degree than to the mind, while possibility of upload and multiple presence (in order of significance) were ascribed to the mind to a significantly higher degree than to the soul. Only the presence in animals was ascribed to both entities to a similar degree (see fig. 6).

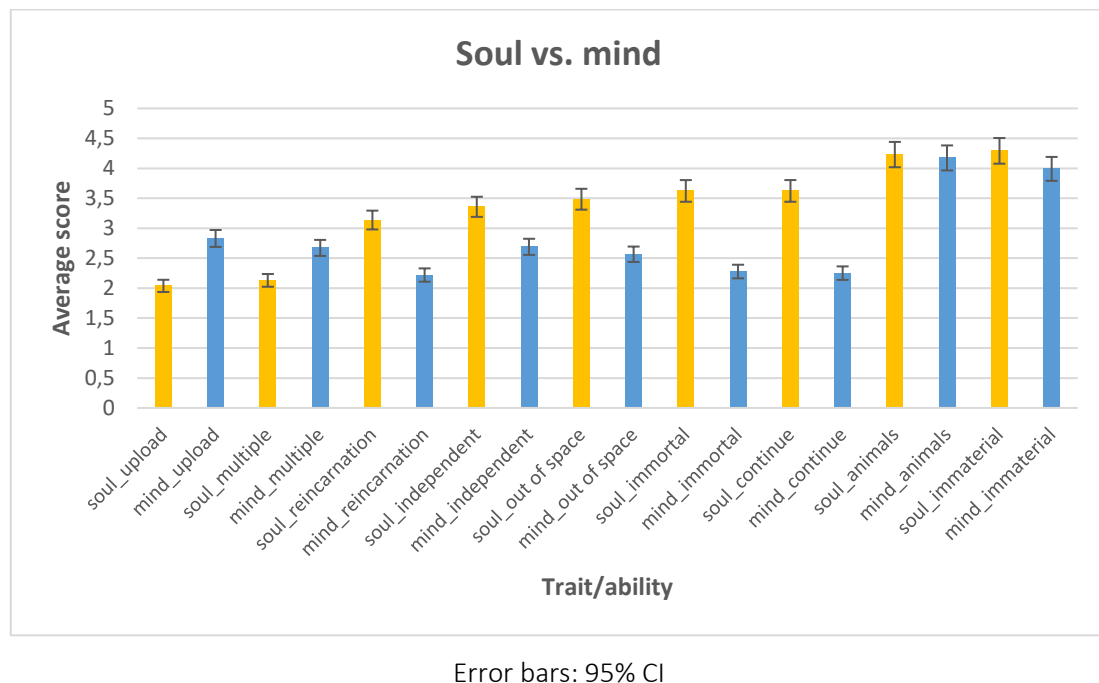


Fig. 6

Mean values of the original scores of the willingness to ascribe different items to either mind or soul (y-axis) as they differ across nine items (x-axis). The scores ascribed to soul are marked in yellow colour, while the scores ascribed to mind are marked in blue. The soul was viewed as more independent and immortal, while the mind was more connected to futuristic scenarios (upload) and presence of multiple (minds) in one body.

²⁷⁸ I performed two-sample two-tailed t-tests, and the results were highly significant for almost all questions ($p < 0.0001$), except for the “animal” question ($p = 0.17$).

Our results suggest that the soul is more connected to the idea of essence, stability, and identity (via its independence and ability to survive death), while mind, on the contrary, is viewed as more suited to enter into situations that might be viewed as threatening identity (such as mind upload or multiple minds in one body). It is thus obvious that these two concepts have different connotations to a nonnegligible degree.

To sum up, our experimental results support the view that folk dualism is more complex than some authors might have originally thought. We can see that it is essential to distinguish between the concept of mind and the concept of soul. The folk concept of mind seems to refer to a mixture of traits across categories, thus the mind seems to be a kind of “interface” between the “essential moral self” traits and all the other “morally neutral” traits – physical and intellectual capacities, or, in other words, between the soul and the brain. Mind is thus viewed as less “transcendent” with respect to the physical world than the soul, probably because it is closely connected to the concept of brain as a source of more “mundane” and morally neutral cognitive capacities.

By referring to a number of empirical studies I have shown that the concept of *essential moral self* and *soul* turn out to be related to a high degree, and together they seem to lie in the core of folk dualism. People show willingness to accept that certain personal traits, including certain aspects of the mind (esp. cognitive and intellectual abilities), can change or cease to exist throughout human life without seriously disrupting personal identity. When it comes to moral traits connected to interpersonal relationships, however, the folk seem to view them as inseparable from the person. They form the essential moral self, the soul, which comes forward as the last untouchable aspect of a human being. Even though people in general might not be willing to endorse such a view explicitly, their tendency to value primarily certain type of traits and their reluctance to ascribe them fully to the physical brain seem to reveal their deep-rooted dualistically coloured concept of a person. Thus, a being worth of the “person status” is not only a free agent capable of escaping causal chain and physical laws but also a being capable of the right moral conduct and the right emotional responses in the context of interpersonal relationships.

We have encountered an example of the same paradox that came forward in the previous chapters. When we are exposed to the scientific facts that reveal human

beings as subjects to physical laws and reduced to their bodies and brains, we tend to realize that something is not right. Even though we rationally accept these truths, we still show reluctance to accept them fully. This is the moment when the two images and the two realms collide, and we are unnaturally forced to choose between them. We can explicitly say that we don't believe in “souls”, but at the same time we are just not comfortable to accept that loving our close ones depends 100% on the physical brain. Similar to the irreducible freedom of a human agent, moral traits that play role in interpersonal relationships seem not to belong to the scientific image of man in the world, and they are not graspable by the language of the physical realm. At the same time, they are so important to us that without them and without the framework capable of grasping them we would not be persons – we would be soulless machines.²⁷⁹

I have already addressed two sub-concepts of the minimal self: the *causal agent* and the *essential self*. Now we need to address the third sub-concept – *conscious subject*. We need to ask what the role of consciousness in the folk concept of self is, and to what extent the philosophical problem of consciousness²⁸⁰ belongs to the folk image of man in the world. In the following chapter, I will try to show that while the “body-soul” dualism described above is a natural component of folk dualism, the “consciousness-brain” dualism is not.

²⁷⁹ In the following chapter I will demonstrate how this point reappears in the studies that thematise robots.

²⁸⁰ Sellars seems to point to this problem when he speaks about the “intrinsic character” of sensations. See Sellars, W., *Science, perception and reality*, p. 36.

4 The Concept of Consciousness

What I choose to call *body-soul* dualism is the above described tendency to view a human being in two modes: as a body/physical object that can be described by the language of science, and as a soul or moral self that can only be graspable by the language of the mental realm and within the manifest framework that enables us to talk about persons. In philosophy of mind, there is yet another type of dualism: *consciousness-brain* dualism, which brings with it the problem of the “explanatory gap” asking how physical events in the physical brain could possibly give rise to (non-physical) phenomenal consciousness.²⁸¹ In this section, I will introduce a current theory that explores the relationship between folk dualism and the philosophical “consciousness-brain” dualism. This theory claims that the explanatory-gap intuition in the philosophy of consciousness belongs to folk dualism and has roots in so-called dual-process cognition. I choose this particular theory because its authors lead very interesting lively debates in the context of experimental philosophy of consciousness. In the second section I will confront this account by introducing a number of other studies focusing on the folk concept of consciousness. I will argue that the real folk dualism does not necessarily include consciousness-brain dualism (or, in other words, philosophically understood problem of consciousness). Even though consciousness has certain role in folk dualism, its logic and its respective explanatory gap is based on something different, something that revealed itself already in the previous chapter: moral and interpersonal dimension.

In the last section of this chapter I will further point to how the folk concept of consciousness is often misunderstood. I will demonstrate how the folk concept differs from the philosophical concept of consciousness referring to Richard Rorty’s, Wilfrid Sellars’s, and John Searle’s criticism of traditional philosophical approaches.

²⁸¹ Another problem related to the hard problem of consciousness is the *meta-problem of consciousness* that asks for explanation of why people feel that there is the hard problem in the first place (even though this is questionable in the case of the folk, as I will show shortly). See e.g. Chalmers, D. (2018). The meta-problem of consciousness. *Journal of Consciousness Studies*, 25(9-10), 6-61.

4.1 Consciousness-brain dualism and the explanatory gap

It is certainly natural (and useful) for humans to think about minds and souls - the carriers of important cognitive and interpersonal tendencies and moral cores, but it is questionable whether thinking about “qualia” and “1st person phenomenal states” could be understood as a natural part of folk dualism. Non-philosophers rarely reflect upon the hard problem of consciousness, while concepts connected to the mind and essentialism about persons seem to be unavoidable.

However, it is possible to claim that the hard problem of consciousness²⁸² is an important aspect of folk dualism. This is what the authors of the following account take as their premise. In their paper on psychological origins of dualism, Brian Fiala, Adam Arico, and Shaun Nichols²⁸³ refer to Bloom to support their claim: “*The rift between consciousness and the physical world is taken to be one central element of folk dualism...*”²⁸⁴ It is apparent that one of the crucial beliefs of folk dualism is that mental states are independent from the body and capable of surviving death, thus in some respect irreducible to physical phenomena.²⁸⁵ The authors presuppose the connection between folk dualism and explanatory gap intuition and go on to develop their main arguments.

The main aim of the authors is to defend the thesis that intuitions about irreducibility of conscious states (explanatory gap) have origins in so-called *dual-*

²⁸² See e.g. Chalmers, D. J. (2010). *The character of consciousness*. New York: Oxford University Press.

²⁸³ Fiala, B., Arico, A., & Nichols, S. (2011). On the psychological origins of dualism: Dual-process cognition and the explanatory gap. In E. Slingerland & M. Collard (Eds.), *Creating consilience: Integrating the sciences and the humanities*. New York: Oxford University Press. Here I refer to the online version retrieved from http://www.u.arizona.edu/~arico/PsychOriginsDualism_final.pdf

²⁸⁴ Ibid., pp. 2f.

²⁸⁵ The belief that mental states are independent from the body and to some extent survive after death seems to be indeed a natural part of folk dualism, as I briefly mentioned in the previous chapter when addressing dualistic tendencies. The question remains, however, to what extent the folk connect this intuition to the hard problem of consciousness as described by philosophers. I will try to show that the typical philosophical view of the consciousness problem has no place in folk dualism.

*process cognition.*²⁸⁶ They try to show that our capacity to attribute conscious states to an entity falls under this special type of cognitive processing: “...*there are two cognitive pathways by which we typically arrive at judgments that something has conscious states (...)* On the one hand, we propose a “low-road” mechanism for conscious-state attributions that has several characteristic System 1 features: it is fast, domain-specific (i.e., it operates on a restricted range of inputs) and automatic (the mechanism is not under conscious control). On the other hand, there are judgments about conscious states that we reach through rational deliberation, theory application, or conscious reasoning; call this pathway to attributions of conscious states “the high road.”²⁸⁷

However, the capacity to attribute conscious states, according to the authors, does not come as a separate cognitive function. Rather, it is inseparable from the capacity to attribute *agency*. We ascribe conscious states only to those entities that we identify as agents (so-called “*AGENCY model*”).²⁸⁸ In order to be identified as an agent, an entity needs to fulfil certain requirements – it has to be represented as having certain features – “*relatively simple, surface-level features, which are members of a restricted set of potential inputs to the low road process. Previous research has identified three features that reliably produce AGENT categorization: that the entity appears to have eyes; that it appears to behave in a contingently interactive manner; and that it displays distinctive (non-inertial) motion trajectories.*”²⁸⁹

²⁸⁶ Dual-process theory claims that “*mental systems fall into two classes. In one class, System 1, we find processes that are quick, automatic, unconscious, associative, heuristic-based, computationally simple, evolutionarily old, domain-specific and non-inferential. In the other class, System 2, we find processes that are relatively slow, controlled, introspectively accessible, rule-based, analytic, computationally demanding, inferential, domain-general, and voluntary.*” (Ibid., p. 3.)

²⁸⁷ Ibid., p. 4.

²⁸⁸ Arico, A., Fiala, B., Goldberg, R. F., & Nichols, S. (2011). The folk psychology of consciousness. *Mind and Language*, 26(3), 327-352.

²⁸⁹ Fiala, B., Arico, A., & Nichols, S., *On the psychological origins of dualism: Dual-process cognition and the explanatory gap*, p. 5.

In one of the studies the authors refer to,²⁹⁰ Johnson, Slaughter, and Carey showed how 12-month old children react to various novel objects via observing their gaze-following behaviour. They presented infants with a brown furry object. The object either had a face composed of simple eyes, nose, and ears (*face condition*), or not (*no-face condition*). In the other pair of conditions the object either interacted by flashing lights and beeping in response to the child's behaviour (*contingent condition*), or remained still and silent (*non-contingent condition*).²⁹¹ When the object had a face, the gaze-following proved to be more intense, and similarly so when the object exhibited contingent behaviour.²⁹² Johnson et al. interpret these effects as proclivity (of both children and adults) to categorize the entity as an *intentional being* in cases when it has a face or displays interactive behaviour. This is also connected to a tendency to attribute mental states to the entity.²⁹³

Fiala, Arico, and Nichols propose that the cognitive process responsible for agency attribution figures also in our intuitions about consciousness: "*In addition to imitation, gaze-following, and reasoning about beliefs and desires, we suggest that agent-categorization also plays a central role in disposing people to regard entities as capable of having conscious experiences.*"²⁹⁴ They further suggest that their theory is empirically testable – people should not have any instant intuitive tendency to ascribe conscious states to objects which do not exhibit the above described required "agency" features such as having eyes and behaving interactively. They should, on the other hand, be prone to attribute agency to the entities that do have these superficial clues.²⁹⁵

The last claim was empirically tested in an experiment in which the authors tested reaction times of respondents who were supposed to decide whether objects presented to them feel pain. They hypothesised that the reaction times will be slower

²⁹⁰ Ibid.

²⁹¹ Johnson, S., Slaughter, V., & Carey, S. (1998). Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. *Developmental Science*, 1(2), 233-238, pp. 234f.

²⁹² Ibid., pp. 236f.

²⁹³ Ibid., p. 237.

²⁹⁴ Fiala, B., Arico, A., & Nichols, S., *On the psychological origins of dualism: Dual-process cognition and the explanatory gap*, p. 6.

²⁹⁵ Ibid., pp. 6f.

when the respondents *deny* conscious state attribution to objects that are typically regarded as agents, while reaction times in non-agent cases will be comparatively faster: *“The idea is that if someone were to overtly respond that entities categorized as AGENTS don’t feel pain (e.g. because they lack appropriate neural structures), this would require overcoming the hypothesized low-road disposition to attribute conscious states to those entities, which would take some extra time.”*²⁹⁶

The participants were asked to respond promptly (Yes or No) to a sequence of Object/Attribution pairs (e.g., *“ant/feels pain”*, or *“feels happy”*). The objects ranged from mammals, birds, insects, and plants to non-living things such as vehicles or natural objects such as rivers or clouds. The subjects were willing to attribute consciousness to objects possessing the relevant features (like mammals, birds, and insects), while their responses were negative when it came to attributing consciousness to inanimate objects typically lacking the right cues. As for the response times, participants were significantly slower in denying consciousness to objects exhibiting superficial agency cues, namely insects. The authors interpret this result as supporting their hypothesis of low-road consciousness attribution: *“...in order to deny that insects have conscious states, subjects had to “override” the low-road output, which explains why reaction times are slower in such cases.”*²⁹⁷

In order to complete the picture of the context within which the intuitive impact of explanatory gap arises the authors also describe the high-road mechanism for attributing conscious states which includes deliberation and inferential reasoning. This can be based on some scientific theory, for example. Knowledge of neural systems and their functioning can lead us to infer that certain organism can feel pain. The pathway that leads us to the final consciousness attribution *“has features typically associated with System 2: processing that is domain general, voluntary and introspectively accessible. The process is domain general in that the inputs are not restricted – evidence can potentially come from anywhere. The process is voluntary because we can control when reasoning starts and stops. And it is introspectively accessible because the steps of the inferential process are typically available to consciousness.”*²⁹⁸

²⁹⁶ Ibid., p. 7.

²⁹⁷ Ibid., p. 7.

²⁹⁸ Ibid., p. 8.

Thus, the high-road mechanism differs from the low-road mechanism in that it is not automatic and unconscious, and it takes more time because of introspectively accessible process of reasoning and considering of available information. The two processes often converge, but they can be in conflict in certain cases. The authors suggest that this happens, for example, when we consider mental life of insects: they have the superficial features that trigger the low-road mechanism leading to conscious states attribution (eyes, certain interactive behaviour), but our knowledge of their limited neural system makes us reconsider the immediate intuition.²⁹⁹ We can already start to see that the authors point to a possibility that a similar conflict plays an important role in our intuitions about the explanatory gap.

Typically, when we look at other people, the low-road mechanism is deployed by the relevant agency cues, and the high-road mechanism leads to agreement with it – thanks to the general knowledge we have about people. The problem arises when we consider the brain alone – a biological mass of brain cells that is the key component of reductionist physicalist theories. If we identify consciousness with certain brain activity pattern of neurons we will infer – via the high-road mechanism – that an entity which displays such and such brain activity pattern is in fact conscious. However, since we consider the person’s brain only, there are no cues present which would trigger the low-road mechanism of consciousness attribution. Hence, we experience tension between the two systems, with one attributing consciousness to the brain and the other remaining silent.³⁰⁰

This is how Fiala, Arico, and Nichols explain our intuitive feeling that physicalist theories of consciousness “leave something out”: *“In place of the harmony between systems that we typically experience when looking at other people (or any other mammal, for that matter), discussions of neurons, neurotransmitters, and so on create a disparity between the two systems, which in turn produces a feeling that the characterization is somehow incomplete. This, we think, is an important part of the explanation for why dualism is so attractive and the explanatory gap is so vexing.”*³⁰¹

²⁹⁹ Ibid., pp. 9f.

³⁰⁰ Ibid., p. 10.

³⁰¹ Ibid., p. 11.

This explanation is plausible also from the perspective of evolutionary (or possibly developmental) theories. The low-road mechanism is sensitive to organisms, and more precisely, to the *outer features* of organisms – the features that are naturally accessible to our senses. Thus, being able to ascribe agency to an organism based on superficial cues might be an adaptation to our environment. There is no reason, then, why we should be able to automatically ascribe consciousness to brains, which are inner bits of organisms and normally hidden from us: “*We (and our ancestors) interacted most often with entire organisms, not neurons in a petri dish. (...) Suborganismic features like neuronal firing patterns never had a chance to shape the mechanism, because they are hidden away behind skin and bone.*”³⁰²

I agree with Fiala, Arico, and Nichols in their claim that there is certain explanatory gap, but I believe it arises within the tension between the description of the physical brain and the description of persons, not between the description of the brain and irreducible aspect of phenomenal consciousness. I also agree that one part of our ability to attribute agency to an entity is also conscious-states attribution. However, I have certain doubts about whether this really addresses the explanatory-gap problem as it is typically discussed in philosophy of consciousness. When the folk attribute conscious states to another being, they are concerned about mental states such as intentions, beliefs, desires, motivations, perceptions, etc. – in short, items belonging to the Davidsonian mental realm, and not qualia as such.

If we ascribe, say, pain to an entity, it can help us predict further behaviour of this entity and adjust our own behaviour (e.g. we can feel that it is morally wrong to hurt a being capable of feeling pain, or we can use this fact to manipulate the being if we see them as our enemy). We need to recognize beings and persons in our environment that are capable of feeling pain or anger in order to adjust our treatment of them accordingly. We do not need to contemplate the irreducibility of their mental states to physical phenomena and the 1st person character and “what-is-it-likeness”³⁰³ of their experiences. We only need to have an appropriate theory of mind in order to categorize them properly and be able to predict further interactions. I believe that laymen actually don’t recognize the philosophical problem of consciousness in the

³⁰² Ibid., p. 20.

³⁰³ Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83(4), 435-450.

same way philosophers do, and in the following section I will turn to relevant literature that provides empirical support for my claim.

4.2 Experimental philosophy of consciousness

It was only quite recently that philosophers of mind turned their attention to actual folk intuitions about consciousness. Joshua Knobe and Jesse Prinz started the debate with their experimental studies based on reported willingness of the folk to ascribe different mental states to group agents.³⁰⁴ Knobe and Prinz believed that they could reveal important aspects of folk beliefs about phenomenal consciousness via observing what kind of mental states folk typically ascribe and don't ascribe to atypical agents – groups.³⁰⁵ They had a presupposition that groups are often viewed as agents, but not as experiencers, thus people should not be keen to ascribe phenomenal states to them.³⁰⁶ The results of their four studies supported this view, and people were indeed significantly less likely to ascribe phenomenal states than non-phenomenal mental states to group agents.³⁰⁷

The main message of the paper is, however, not as much addressing the concrete form of the folk concept of phenomenal consciousness, but rather the way folk *use information* about whether the entity is or is not phenomenally conscious. Knobe and Prinz put their view in contrast to functionalist “grand view” theories of folk psychology that understand mental states as certain tools helping us predict and explain behaviour of agents. They propose that the case of ascription of phenomenal mental states serves rather as a compass for subsequent *moral judgements*. For example, if we believe that an entity is capable of experiencing pain, the right moral thing to do for us would be to avoid hurting it.³⁰⁸ Knobe and Prinz support their view

³⁰⁴ Knobe, J., & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, 7(1), 67-83. Here I refer to the advance online publication retrieved from <https://cpb-us-w2.wpmucdn.com/campuspress.yale.edu/dist/3/1454/files/2016/02/Consciousness-28f03o2.pdf>

³⁰⁵ Ibid., p. 6.

³⁰⁶ Ibid., p. 13.

³⁰⁷ Ibid., pp. 7-12.

³⁰⁸ Ibid., pp. 13-16.

by two further studies and show that people indeed use information about phenomenal consciousness to make moral rather than functional judgements.³⁰⁹ In one of the studies, they asked two groups of participants why a person responsible for taking care of fish would be curious about their memory ability (group 1) or about their ability to feel pain (group 2). As Knobe and Prinz expected, almost all subjects in the memory condition mentioned solely explaining, predicting, and controlling behaviour, while all subjects in the consciousness condition brought out moral considerations.³¹⁰

Work of Knobe and Prinz was followed by numerous critical replies that attempted to explain their results in alternative ways. Justin Sytsma provides an overview of this debate³¹¹ and addresses the shortcomings of the vignette presented by Knobe and Prinz. First, he refers to Adam Arico's critic³¹² of the contextual disbalance between the phenomenal and non-phenomenal versions of the scenarios.³¹³ Further he mentions a reply by Phelan et al.³¹⁴ who demonstrate the possibility that people assign mental states to group agents by means of perceiving them in a distributivist fashion (as mental states of the particular members of the group in question) and not as real agents in the usual sense, thus putting the conclusions of Knobe and Prinz in doubt.³¹⁵ Another critical contribution, this time explicitly suggesting that the folk don't have the alleged concept of phenomenal consciousness, comes from Justin Sytsma and Edouard Machery.³¹⁶ Via their own experiments, they put forward a suggestion that lay people (in contrast to

³⁰⁹ Ibid., pp. 16-19.

³¹⁰ Ibid., pp. 17ff.

³¹¹ Sytsma, J. (2014). Attributions of consciousness. *WIREs Cognitive Science*, 5(6), 635-648.

³¹² Arico, A. (2010). Folk psychology, consciousness, and context effects. *Review of Philosophy and Psychology*, 1(3), 371-393.

³¹³ Sytsma, J., *Attributions of consciousness*, pp. 639f.

³¹⁴ Phelan, M., Arico, A., & Nichols, S. (2013). Thinking things and feeling things: on an alleged discontinuity in the folk metaphysics of mind. *Phenomenology and the Cognitive Sciences*, 12(4), 703-725.

³¹⁵ Sytsma, J., *Attributions of consciousness*, pp. 640f.

³¹⁶ Sytsma, J., & Machery, E. (2010). Two conceptions of subjective experience. *Philosophical Studies*, 151(2), 299-327.

philosophers) don't ascribe mental states based on their phenomenality, but rather based on whether they have hedonic value or not (lay people in their experiment were willing to ascribe seeing, but not feeling pain to a robot).³¹⁷ Sytsma, however, refuses this explanation in the end and prefers so-called “*naïve account*,” saying that “*lay people tend to view both colours and pains not as phenomenal qualities, but as mind independent qualities of objects outside the brain/mind.*”³¹⁸

When developing his theory of the *naïve account* of the folk concept of phenomenal consciousness, Sytsma argues against the tendency of professional philosophers to ascribe belief in qualia to all people in general.³¹⁹ Sytsma demonstrates the key error by reviewing the case of philosopher of mind Daniel Dennett, who believes that current scientific view of perceptual experience is widespread.³²⁰ Dennett argues that there “*seem to be qualia*” because science has uncovered the truth about the existence of colours and other phenomenal qualities: they do not exist in the outer world, thus they must exist inside our heads.³²¹ Sytsma correctly notes that there is no good reason to consider any view to be widespread just because it is in accord with the current scientific account. As an example, he refers to studies that clearly demonstrate that despite the scientifically accepted intromissionist view of vision, many people still have a tendency to believe that “*something goes out of the eyes*” in the process of vision (so-called “extramissions”).³²² This seems to remain the case even after the subjects undergo relevant education.³²³

³¹⁷ Sytsma, J., *Attributions of consciousness*, pp. 641ff.

³¹⁸ Ibid., p. 643.

³¹⁹ Sytsma, J. (2010). Dennett's theory of the folk theory of consciousness. *Journal of Consciousness Studies*, 17(3-4), 107-130. Here I refer to the advance online version retrieved from http://philsci-archive.pitt.edu/5141/1/Dennett%27s_Theory_of_the_Folk_Theory_of_Consciousness.pdf

³²⁰ Ibid., pp. 2f.

³²¹ Ibid., p. 4. (Sytsma quotes Dennett, D. C., *Consciousness explained*, p. 372.)

³²² Sytsma, J., Dennett's theory of the folk theory of consciousness, p. 5.

³²³ Gregg, V. R., Winer, G. A., Cottrell, J. E., Hedman, K. E., & Fournier, J. S. (2001). The persistence of a misconception about vision after educational interventions. *Psychonomic Bulletin & Review*, 8(3), 622-626.

Sytsma further points out that Dennett puts folk theory of consciousness in close relation to certain philosophical non-reductionist views of consciousness, such as Thomas Nagel's "what is it like"³²⁴ and David Chalmers's "hard problem of consciousness"³²⁵ notions.³²⁶ In order to demonstrate the fallacy hidden in this approach, Sytsma not only reviews current debate on folk concept of phenomenal consciousness, but he also introduces new experiments addressing folk view of colours and pains.³²⁷ Respondents (adult Americans) with no previous training in philosophy or psychology supported Sytsma's hypotheses: majority of them leaned towards the view that colours are properties of the objects outside of us, that they are independent from the mind, and that the spectrum inversion (that someone could see a tomato appearing red to me as blue) is impossible.³²⁸ The same pattern of results appeared also after Sytsma asked his subjects about pains: according to the majority of respondents, pain is a quality located not in our mind, but in the hurt body parts,³²⁹ and they even agreed that unfelt pains were possible.³³⁰ Vignette about shared pains proved once again that lay people lean towards the naïve notion of pains.³³¹

Sytsma sums up that *"the naïve view appears to be the majority view despite the fact that it is a minority view in philosophy. (...) In slogan form, the folk do not treat the red as being in the head and they do not treat the pain as being in the brain."*³³² Number of repeated experiments managed to seriously deconstruct Dennett's confident notion of folk psychology of phenomenal consciousness.

Results and arguments of Sytsma and Machery might not be perfectly convincing for everyone. Anthony Peressini agrees with them to a certain extent (namely that folk don't have a proper concept of qualia) but suggests where Sytsma

³²⁴ Nagel, T., *What is it like to be a bat?*

³²⁵ Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

³²⁶ Sytsma, J., *Dennett's theory of the folk theory of consciousness*, p. 10.

³²⁷ *Ibid.*, pp. 16-26.

³²⁸ *Ibid.*, pp. 17f.

³²⁹ *Ibid.*, pp. 18ff.

³³⁰ *Ibid.*, pp. 20ff.

³³¹ *Ibid.*, pp. 23ff.

³³² *Ibid.*, p. 25.

and Machery go too far.³³³ He believes that “*phenomenological component*” should not be completely refused as a possible aspect of the folk concept of subjective experience and provides more direct experiments to prove his point.³³⁴ Peressini attempted to aim at folk willingness to attribute “*being an experiencer*” to certain different entities and at the folk concept of qualia.³³⁵ Despite his respondents were generally unwilling to ascribe the status of an experiencer to robots,³³⁶ his results were put in doubt due to a serious priming problem:³³⁷ he first introduced an explicit concept of an experiencer giving concrete examples (human beings vs. thermostats) and even used expressions like “inner life”, “inside our minds”, etc.³³⁸ The point of indirect questions, like in Sytsma and Machery, is precisely to avoid this kind of priming, despite certain disadvantages of this model. In my opinion, the conclusions of Sytsma and Machery remain untouched by studies that don’t avoid the priming problem.

When we return to the agency model of Fiala, Arico, and Nichols, the above-mentioned and other similar studies seem to have certain implications for their notion. Namely, the fact that respondents were willing to ascribe the ability to see or smell something, but not the ability to feel pain or experience emotions³³⁹ to a robot, might partially go against the agency model, for this model predicts ascription of the whole spectrum of mental states without such specific exceptions.³⁴⁰

³³³ Peressini, A. (2013). Blurring two conceptions of subjective experience: Folk versus philosophical phenomenality. *Philosophical Psychology*. Advance online publication. DOI: 10.1080/09515089.2013.793150.

³³⁴ Ibid., p. 6.

³³⁵ Ibid., p. 7.

³³⁶ Ibid., pp. 8-12; 17.

³³⁷ Sytsma, J. (2014), *Attributions of consciousness*, p. 644.

³³⁸ Peressini, A., *Blurring two conceptions of subjective experience: Folk versus philosophical phenomenality*, pp. 7f.

³³⁹ Heubner, B. (2010). Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? *Phenomenology and the Cognitive Sciences*, 9(1), 133–155.

³⁴⁰ Fiala, B., Arico, A., & Nichols, S. (2014). You Robot. In E. Machery & E. O’Neill, (Eds.), *Current controversies in experimental philosophy* (pp. 31-47). Abingdon: Routledge, p. 35.

However, the authors see the answer to reconciling the results with the agency model in its very cornerstone – the dual process cognition theory. It might be simply the case that the subjects in these studies had just enough time to employ high-road reflection. The widespread “knowledge” that robots don’t have emotions and cannot feel pain came to the surface and outweighed the low-road tendency to ascribe full agenthood to the robot based on initial superficial clues.³⁴¹ This idea was supported by new experimental studies, in which Fiala, Arico, and Nichols provided respondents with a wider range of possibilities how to express what the robot does in a particular task in the vignette when manipulating boxes of a particular colour (not only “saw”, but also “detected”, “identified”, and “located”). They found that people were less likely to select “saw” or “know” in the case of a robot, but to some extent even in the human case, when given more possibilities how to express what was going on in the vignettes. This clearly indicates that there must be “*a fair amount of metacognition going on in these cases.*”³⁴²

Fiala, Arico, and Nichols go on to support their thesis by listing the evidence of our low-road tendency to see robots as agents and attribute to them mental states that seem to be omitted as a result of the high-road reflexion. According to their agency model, categorisation of an entity as an agent is based on three main features: *distinctive motion trajectories, presence of face and contingent interaction.*³⁴³ They provide evidence showing that many robots or computer systems are capable of fulfilling these.³⁴⁴ In cases when the robot fulfils all the basic agency features, the human responses are especially interesting. The authors refer³⁴⁵ to an experiment by Bartneck, Van Der Hoek, Mubin, and Al Mahmud³⁴⁶ in which they programmed an iCat robot to be able to speak, mimic human expressions of basic emotions, and cooperate with human subjects in playing a game (Mastermind) against a computer.

³⁴¹ Ibid., p. 37.

³⁴² Ibid., pp. 37-40.

³⁴³ Ibid., p. 40.

³⁴⁴ Ibid., pp. 40-43.

³⁴⁵ Ibid., p. 43.

³⁴⁶ Bartneck, C., Van Der Hoek, M., Mubin, O., & Al Mahmud, A. (2007, March). “Daisy, daisy, give me your answer do!” Switching off a robot. *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction, Washington DC* (pp. 217–222). New York, NY: ACM.

The robot manifested either high or low intelligence (smart vs. stupid condition) and was either polite or unmannerly (high vs. low agreeableness condition). The participants were supposed to turn the robot off after the game (being informed that this will delete robot's memory and personality) while the robot expressed unwillingness to be switched off.³⁴⁷ The respondents were not able to switch the robot off without hesitation. Politeness of the robot prolonged the hesitation to more than double in the smart condition, and in the case when the iCat was both polite and smart the participants took three times longer to turn it off than in the negative case (stupid and impolite).³⁴⁸ Another example³⁴⁹ was taken from the work of Bartneck and Hu³⁵⁰ and their experiment with a "Crawlig Microbug" robot. Participants were supposed to "kill" the robot using a hammer after interacting with it for three minutes, and they were told that this act was necessary for the success of the study.³⁵¹ Judging by the remarks some participants provided after the experiment, they felt it was morally wrong to destroy the robot.³⁵²

Sytsma believes that with all this being said, the agency model of Fiala, Arico, and Nichols is in accord with the findings of Sytsma and Machery.³⁵³ The only tension concerns the role of phenomenal mental states in the judgements that employ high-road cognitive processes. Sytsma stands behind the results that showed reluctance of participants to ascribe *certain specific* mental traits to the robot, such as feeling pain or emotions, as opposed to mental states in the wide sense, as predicted by the theory of Fiala, Arico, and Nichols.³⁵⁴ Some phenomenal mental states (such as seeing or smelling) do not fall into the category of states that would be considered as inappropriate to ascribe to robots even after high-road reflexion, thus the

³⁴⁷ Ibid., pp. 218f.

³⁴⁸ Ibid., pp. 220f.

³⁴⁹ Fiala, B., Arico, A., & Nichols, S., *You Robot*, pp. 43f.

³⁵⁰ Bartneck, C., & Hu, J. (2008). Exploring the abuse of robots. *Interaction Studies*, 9(3), 415–433.

³⁵¹ Ibid., p. 422.

³⁵² Ibid., p. 426.

³⁵³ Sytsma, J. (2014). The robots of the dawn of experimental philosophy of mind. In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy* (pp. 48-64). Abingdon: Routledge, p. 53.

³⁵⁴ Ibid., p. 62.

phenomenality of the mental states *as such* cannot play the central role in the agency model as Fiala, Arico, and Nichols suggested in their work on dual-process cognition.³⁵⁵ Only a specific subset of phenomenal mental states seems to be essential for the narrower and more “exclusive” concept of an agent, namely those that play a role in our *moral considerations*.³⁵⁶

Despite their failure to properly capture the folk concept of phenomenal consciousness, Knobe and Prinz were on the right track when they suggested that moral considerations play a crucial role in attribution of phenomenal mental states.³⁵⁷ Bryce Heubner comes to a similar conclusion: “*I close by arguing that disputes over the philosophical notion of ‘phenomenal consciousness’ are misguided and that they fail to capture the important role of moral consideration in determining whether an entity is a locus of subjective experience.*”³⁵⁸ In his experiments, participants were significantly less likely to ascribe pain³⁵⁹ and emotion (feeling happy)³⁶⁰ to a robot than to a human being, while ascribing belief to a robot was unproblematic for the participants in both vignettes. Heubner suggests that by ascribing the ability to feel

³⁵⁵ Fiala, B., Arico, A., & Nichols, S., *On the psychological origins of dualism: Dual-process cognition and the explanatory gap*.

³⁵⁶ Our own interview study (Kopecký, R., & Jirout Košová, M., not yet published) with 209 Czech children and teenagers (52.15% F, average age 11.05 years, age range 6-17) is in accord with this view. Participants were supposed to ascribe different capabilities/features (on a 6-point Likert scale: 0-5 points) to a dog, simple mathematical robot, and an android who perfectly wields human language and is capable of sophisticated communication with humans (children in the pilot study ascribed all these abilities fully to a human being, so we did not include human again in the main study). The results showed that participants were willing to ascribe to the android the ability to see or hear, the ability to think (4 to 5 points in average), the ability to decide freely for itself (between 3 and 4 points), to a considerably lesser extent the ability to feel happy and to realise who they are (self-reflexion) (slightly over 3 points), and they were generally not willing to ascribe life and soul (slightly over 1 point) to the android.

³⁵⁷ Knobe, J., & Prinz, J., *Intuitions about consciousness: Experimental studies*, pp. 17ff.

³⁵⁸ Heubner, B., *Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies?*, p. 134.

³⁵⁹ *Ibid.*, pp. 137-141.

³⁶⁰ *Ibid.*, pp. 141-144.

pain and happiness to an entity, we acknowledge the entity as caring for “*how things turn out for her*”, and thus the entity becomes “a subject for moral concern.”³⁶¹ He further stresses that there simply is not a “*single uniform strategy that is adopted by all people in ascribing mental states.*” In order to understand these issues, we need to reflect upon “*the relationship between our ascriptions of mental states and the structure of our moral psychology.*”³⁶² Heubner, referring to Haslam et al.,³⁶³ introduces the distinction between *agency* and *personhood* strategies of ascribing mental states. While the *agency strategy* is focused on rational goal-oriented behaviour, *personhood strategy* is concerned with moral considerations (believing that the entity cares for how things turn out for her). The strategies go mostly hand in hand, for we usually interact with agents who also happen to be persons, but they can easily diverge when we encounter more challenging cases.³⁶⁴ This approach is akin e.g. to Dennett’s distinction between the *intentional* and the *personal* stance, in which the personal stance presupposes the intentional stance while having an additional “*moral commitment*”.³⁶⁵

Especially strong support for these ideas comes from research that returns us to the problem of *connection between consciousness and free will*. Nahmias, Allen, and Loveall attempt to find this connection via their own studies.³⁶⁶ I have suggested in the second chapter that consciousness plays an important role in the folk concept of free will and pointed to the idea of *conscious will* that may facilitate the ultimate unity of the self and the “transcendent” aspect of free will – the folk belief that human agents escape the physical chain of deterministic connection between causes and effects. But why does free will has to be conscious? Nahmias et al. rightly

³⁶¹ Ibid., pp. 148f.

³⁶² Ibid., p. 150.

³⁶³ Haslam, N., Kashima, Y., Loughnan, S., Shi, J., & Suitner, C. (2008). Subhuman, inhuman, and superhuman: contrasting humans and nonhumans in three cultures. *Social Cognition*, 26(2), 248–258.

³⁶⁴ Heubner, B., *Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies?*, pp.150f

³⁶⁵ Dennett, D. C. (1981). *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: The MIT Press, p. 240.

³⁶⁶ Nahmias, E., Allen, C. H., & Loveall, B., *When do robots have free will? Exploring the relationships between (attributions of) consciousness and free will*.

conclude that the connection between consciousness and free will is “*under-analyzed*” probably because it seems so obvious. Agenthood is, after all, always connected to conscious beings in our everyday lives. This is why the studies introducing all kinds of unintuitive beings (e.g. complex humanoid robots) are so important, because they problematize the apparently obvious connection.³⁶⁷

Nahmias et al. gained inspiration from the studies of Joshua Shepherd who managed to reveal indices pointing to how consciousness relates to free will. His respondents tended to consider robots in his vignettes as morally responsible agents possessing free will, provided that they were described as conscious (capable of seeing colours, but also of feeling pain and emotions, and able to consciously deliberate). Unconscious versions of the robots were much less likely to receive ascriptions of free will and moral responsibility.³⁶⁸

Nahmias et al. sensed the connection between Shepherd’s results and P. F. Strawson’s views concerning the role of reactive attitudes in our notion of freedom.³⁶⁹ They accept Strawson’s suggestion that freedom and moral responsibility in agents is tied to the fact that they are considered to be appropriate “*targets*” of the reactive emotions such as “*indignation, gratitude, and guilt, that we express in interpersonal relationships*”. In order to qualify for this status, agents have to be capable of conscious experiencing of “*certain moral emotions*”, and they have to express them while they act. Thus, the connection between consciousness and freedom appears to dwell in the ability to experience certain specific kind of emotions – the Strawsonian reactive emotions.³⁷⁰

Apart from the Strawsonian notion of reactive emotions, Nahmias et al. refer to views that we encountered already in Heubner. They are concerned with the “*ability to care about what motivates us*”. Caring for something means that the agent is bound to feel different emotions (e.g. satisfaction, joy, disappointment, sadness,

³⁶⁷ Ibid., p. 59.

³⁶⁸ Shepherd, J. (2015). Consciousness, free will, and moral responsibility: Taking the folk seriously. *Philosophical Psychology*, 28(7), 929-946, pp. 939-940.

³⁶⁹ Strawson, F. P. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48, 1-25.

³⁷⁰ Nahmias, E., Allen, C. H., & Loveall, B., *When do robots have free will? Exploring the relationships between (attributions of) consciousness and free will*, p. 64.

etc.) depending on how things turn out for her.³⁷¹ According to this picture, free responsible agent would be someone capable of feeling reactive attitudes as well as emotions tied to the outcomes concerning what she cares about.³⁷²

With this notion in mind, Nahmias et al. proceed with their own online studies on (around 300) American undergraduates, using humanoid robots in the vignettes. In the first (between-groups) study, behaviour of humanoid robots (and humans in the control condition) was described, and the subjects were asked to respond on a 7-point scale to what extent they would ascribe to them “*free will, moral responsibility, basic emotions, Strawsonian emotions, conscious sensation, reason and reflection, and theory of mind.*” The behaviours described in the vignettes are typically connected to conscious experiences in the case of human agents.³⁷³ Even though the robots expressed the same behaviour as humans, the respondents attributed to them less free will, less moral responsibility, and lower ability to experience all kinds of emotions and sensations. The same applied to cognitive abilities and the theory of mind. In general, participants ascribed the whole range of capabilities to robots around the midpoint in average, thus they were not quite sure what to think about the robots, while they were confident about humans.³⁷⁴ However, when Nahmias et al. compared the responses of those participants who ascribed conscious experiences to robots with the results of the other group, they found out that together with ascription of consciousness, the participants ascribed also the rest of the abilities to the robots. This information was important for the design of the subsequent study.³⁷⁵

In the second between-groups study, they introduced the consciousness manipulation – in one condition robots were described as conscious, in the other they were described as non-conscious. Respondents were informed about the mental states and capacities of the robots in the vignettes (that they do/don’t see colours, feel pain, and experience emotions).³⁷⁶ Participants attributed more free will to conscious

³⁷¹ Ibid., pp. 64f.

³⁷² Ibid., pp. 66f.

³⁷³ Ibid., p. 68.

³⁷⁴ Ibid., pp. 69ff.

³⁷⁵ Ibid., p. 72.

³⁷⁶ Ibid., pp. 72f.

robots, but they did not judge them as morally responsible. They were also not willing to ascribe the ability to experience emotions and sensations to non-conscious robots, but the cognitive abilities and theory of mind remained unaffected by the consciousness condition.³⁷⁷

Nahmias et al. were further looking at the details of the relationship between consciousness manipulation and free will (“*potential mediators for the effect of consciousness on free will attributions*”). They found out that emotions are the key: “...*the extent to which people judged the robots as able to experience Strawsonian and Basic Emotions fully mediated the relationship between the consciousness manipulation and people’s attributions of Free Will.*” This supports precisely the point about the importance of *specific subset* of conscious states that I have been trying to demonstrate in this chapter: “*In other words, the ability to feel emotions and have things actually matter to the individual is important in Free Will attributions, yet the ability to have conscious sensations (e.g., the ability to experience sounds or smells) specifically plays no significant role.*”³⁷⁸ The authors conclude that their results support the possible view that the connection between consciousness and free will dwells precisely in the fact that free moral agents have to *care* about the outcomes of their actions and *interpersonal interactions* and feel *moral emotions* tied to them. The rationality, intelligence, and complexity of their behaviour does not play the most essential role here.³⁷⁹

I believe that this illustrates where the theory of Fiala, Arico, and Nichols shows as insufficient: they are not able to explain the cases in which people ascribe certain mental states (e.g. seeing colours) to entities that exhibit agency features but at the same time deny them (as a result of high-road reflexion) the status of moral entities (because they are reluctant to ascribe pain or emotions to them). These are the cases that happen to reveal to us that the concept phenomenal consciousness is not strictly bound to the folk concept of agent as Fiala, Arico, and Nichols define it. Along the lines with what Heubner and Nahmias et al. suggest, it is apposite to introduce a more refined concept of an agent, namely realising that an agent can be viewed as a rational, goal-oriented being capable of sense perception, but also a

³⁷⁷ Ibid., pp. 74f.

³⁷⁸ Ibid., pp. 75f.

³⁷⁹ Ibid., p. 77.

sentient being capable of feeling pain and emotions and requiring to be categorised as a moral entity.³⁸⁰ Lay people seem to be perfectly fine with ascribing the first type of agency to a robot even via high-road reflexion (together with certain types of phenomenal mental states), but they are hesitant when it comes to the question of employment of moral considerations. It is their role *in interpersonal moral conduct*, and not the *phenomenality* of the mental states that is the criterion that determines this categorisation.

Fiala, Arico, and Nichols still tell us something important about folk intuitions. We indeed feel puzzled when we are told that the mental states we and other agents have are rooted in brain activity.³⁸¹ We have a specific idea of agents, which is connected to the outer agency cues that deploy the low-road mechanism of agency attribution and the further idea that each person has a mind (as a rational agent) and a self (as a person falling under moral considerations). We are not wired to react intuitively (via low-road) to brains, since we almost never encounter inner parts of agents in our everyday social lives. We are adapted to our natural environment, which is made of agents and *selves* closed in moving, interacting bodies with faces. Thanks to their complexity we need to approach agents in a radically different manner than we approach material bodies, hence we have this specific concept.

This is where the authors should probably stop. The explanatory gap arises from our folk body-soul (or body-essential moral self) dualism and sensitivity to certain specific cues in our environment. Once the high-road mechanism leads us to acknowledge that the self with all its moral connotations is nothing over and above the physical body and brain activity, we naturally feel that something is not quite right. We are forced to throw away our manifest image, our mental realm, and overrule our deeply rooted dualism. Irreducible subjective and qualitative properties

³⁸⁰ Weisman, Dweck, and Markman also point to this possibility while defending their “*body-mind-heart*” framework of the people’s categorisation of mental life. (Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people’s conceptions of mental life. *Proceedings of the National Academy of Sciences of the United States of America*, 114(43), 11374-11379, pp. 11377f.)

Interestingly, the factor they call “heart” seems to correspond to the folk concept of soul (p. 11375).

³⁸¹ Bloom, P., *Descartes’ baby*, p. 170.

of conscious experience do not need to enter into this altogether. The case with robots is special and therefore very illustrative in that people don't seem to be willing to ascribe *moral selves* ("souls") to robots and thus don't feel perplexed that these purely rational and goal-oriented agents could be reduced to electrical circuits and chips, together with their "phenomenal" mental states such as seeing and hearing.

I believe it is not right to claim that phenomenal concepts are *always* deployed together with the low-road mechanism of agency attribution. What is deployed are certain concepts we have about minds and persons – a range of mental states we are able to ascribe to an agent. These concepts do not have to include the aspect of irreducible subjective properties. This conceptual distinction comes only after deeper philosophical reflexion and analysis. Folk dualism does not have to reflect upon these dimensions of mental life. Only philosophical brain-consciousness dualism includes this additional step. What is more, some might claim that there is no evolutionary reason why humans should be capable of solving problems like the hard problem of consciousness,³⁸² so there is probably no reason why lay people should be generally prone to acknowledge this problem in the first place.

Explanatory gap stemming from folk dualism is not only about the body and the mind. As I concluded in the previous chapter, the main ingredient that makes reductionist theories so vexing is the concept of the *essential moral self* or the *soul*. There is a disharmony between the low-road mechanism that ascribes moral selves to persons and the high-road mechanism by which we come to theories that all that constitutes a person is fully grounded in the physical body and the brain. We are essentialists about the self and understand persons in completely different terms than physical objects. We never had a chance to see the connection between brains and souls in our evolutionary developmental history, and that is why reductionist theories leave us struggling with our deepest intuitions.

Explanatory gap in philosophy of consciousness, on the other hand, has a completely different logic: it is based on the feeling that 3rd person explanations conveyed in objectivist scientific language cannot account for 1st person qualities of subjective experience. What is more, it is rather difficult to grasp these concepts properly – they are *philosophical* concepts, not *folk* concepts. If lay people have a

³⁸² See e.g. Chomsky, N. (2009). The mysteries of nature: How deeply hidden? *The Journal of Philosophy*, 106(4), 167-200.

feeling that their mental lives are part of something that surpasses the body, it is mainly because of their essentialist moral concept of the self. Another reason might be our inability to imagine cessation of our own mental lives – so-called “simulation constraint”.³⁸³ The irreducible subjectivity of conscious states probably plays certain role in these intuitions (as in the concept of *conscious will* that seems to be the basis for the ultimate unity and causal transcendence of the self), but it is, I believe, unconscious and mostly not reflected in the sense in which philosophers usually reflect this problem. Folk simply don’t seem to acknowledge the hard problem of consciousness.³⁸⁴

More research needs to be done to properly explore folk dualism and all the concepts that construct it. Anyway, at this point I prefer to lean towards the account that is more in accord with evolutionary reasoning. It seems to be more adaptive for humans to be natural body-soul dualists than consciousness-brain dualists. The concepts of philosophy of consciousness not only require reflexion and philosophical training, but they also seem unnecessary for our orientation in the world of agents. The explanatory gap that divides the deep moral self – the irreducible core of each human being on the one side, and the attempts by scientific theories to reduce the soul to mere functions of the physical brain on the other side, is more likely to be deeply rooted in us.

³⁸³ See Bering, J. M., *The folk psychology of souls*, p. 455.

³⁸⁴ No doubt further research is needed in the area of the long unaddressed folk concept of consciousness. Rodrigo Díaz from the University of Bern has recently conducted studies in order to test whether the folk have intuitions concerning the hard problem of consciousness and what factors drive these intuitions. (Díaz, R. (in press). Do people think consciousness poses a hard problem? Empirical evidence on the meta-problem of consciousness. *Journal of Consciousness Studies*. Here I refer to the advance online publication retrieved from <https://philarchive.org/archive/DAZDPT>)

His results suggest that “(1) *problem intuitions are not widespread, and (2) when they arise, they do so because of factors that are unrelated to the nature of consciousness. This suggests that consciousness is, after all, not so problematic.*” (p. 1). When the problem intuitions appeared, their presence was explained by scepticism about the completeness and quality of science (*consciousness-independent factors*), and not by the respondents’ tendency to engage in inward thinking and reflexion (*consciousness-related factors*) (pp. 15ff).

We are social beings who spend their whole lives in an immensely intricate social environment and are capable of forming close interpersonal bonds. It is not the irreducibility of the 1st person point of view and qualia, but the moral dimension of the being, that creates the real (non-philosophical) explanatory gap. Being a subject of experience is indeed an important aspect of the minimal self, but the experience that plays role here is not primarily defined by its phenomenality. Instead, the aspect of experience that matters is one that has to do with moral emotions and caring for how things turn out for us and other sentient beings with whom we enter into interpersonal relationships. Just as Paul Bloom's little son Max and respondents from our own online questionnaire study on dualism believe, the brain sees and thinks, but it is just too weird to say that the brain loves somebody. The real abyss dwells between the world of purely scientific descriptions and the world of the essential moral selves.

4.3 How philosophers and the folk see consciousness

In this section I would like to address the *philosophical concept of phenomenal consciousness* in more detail and point to some interesting ideas that were introduced long before the ascent of experimental philosophy and yet illustrate very aptly what x-phi has revealed about the folk concept of phenomenal consciousness.

The whole misunderstanding about the problem of consciousness arises in the moment when some philosophers of consciousness distort the essential *subjectivity* of the 1st person view that non-reductive theories of consciousness try to point to.³⁸⁵ They tend to “objectify” the inner space of our minds by using certain “models” or “metaphors”³⁸⁶ of what is going on in people's heads and talk about the mind as about some “inner space” occupied by uncountable entities called thoughts, beliefs, perceptions, ideas, memories, qualia, etc.³⁸⁷

³⁸⁵ I address this problem in my master's thesis: Košová, M., *Modern Theories of Consciousness and the Elusiveness of Subjectivity*.

³⁸⁶ Consider Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

³⁸⁷ Certain metaphors about the mind are natural for the folk (as e.g. Lakoff and Johnson show in their book), but the extent and consistency of the idea that there are certain mental

There is still plenty of confusion when it comes to the question of consciousness and its place in the physical world. Theories that claim identity of conscious states or sensations with neurological processes³⁸⁸ seem deeply unsatisfactory to many philosophers who cling to an intuition that there is a *hard problem of consciousness*.³⁸⁹ However, it seems equally absurd to seek refuge in dualistic theories of consciousness viewing sensations as parts of special immaterial substance, some sort of “ghost stuff”. Where should we turn, if we just can’t agree with the identity theories and, on the other hand, find the idea of “ghost stuff” long surpassed?

In what follows I attempt to introduce a possible answer by analysing and confronting positions of Richard Rorty, Wilfrid Sellars, and John Searle, who uncover a very interesting possibility - that our problems with capturing conscious states might be rooted in our misleading and distortive conceptualization of them. This will help us to further determine the role consciousness plays (and doesn’t play) in the folk concept of person. Based on the experimental studies introduced in the previous section we can start to see that it is the philosophers (and not the folk) who bring the main confusion on the scene.

Richard Rorty claims that the root of the whole mind-body problem is our unreflective imprisonment in a certain specific language game which influences our concept of the mental.³⁹⁰ Neo-dualists might assert that the phenomenal is non-physical and support this conclusion by the claim that knowing physical properties through-and-through doesn’t entail knowing what it is like to be the being described by those properties – we don’t gain the knowledge of “*how it feels*”. According to Rorty, however, this doesn’t entitle us to infer a gap between the ontological status of the referents of the physiological and the phenomenological terms.³⁹¹

entities *inside our heads* (especially qualia) shows as much weaker in the case of laymen, as the empirical studies I referred to seem to demonstrate.

³⁸⁸ See Smart, J. J. C. (1959). Sensations and brain processes. *The Philosophical Review*, 68(2), Ithaca, NY: Cornell University, 141-156.

³⁸⁹ See Chalmers, D. J., *The character of consciousness*.

³⁹⁰ Rorty, R. M. (1979). *Philosophy and the mirror of nature*. Princeton: Princeton University Press, p. 22.

³⁹¹ *Ibid.*, pp. 28f.

Neo-dualist can still support his case by pointing out that when it comes to phenomenological properties, *“there is no appearance-reality distinction. This amounts to defining a physical property as one which anybody could be mistaken in attributing to something, and a phenomenal property as one which a certain person cannot be mistaken about.”*³⁹² Here Rorty asks the crucial question: *“But why should this epistemic distinction reflect an ontological distinction? Why should the epistemic privilege we all have of being incorrigible about how things seem to us reflect a distinction between two realms of being?”*³⁹³ If the questioned philosopher attempts to answer this question, he’ll get into a slippery talk in *“old-fashioned Cartesian dualist”* style, since he *“stopped talking about pains as states of people or properties predicated of people and started talking about pains as particulars, a special sort of particular whose nature is exhausted by a single property. Of what could such particular be made, save mind-stuff?”*³⁹⁴

Rorty provides his diagnosis as to why this happens: the neo-dualist simply hypostatizes a universal property of a raw feel (e.g. pain) into a special kind of *“particular”*. He is *“no longer talking about how people feel but about feelings as little self-subsistent entities, floating free of people in the way in which universals float free of the instantiations.”*³⁹⁵

By this Rorty provided a clue as to why *“we think of the phenomenal as immaterial”*. Mentioning Ryle, he says that *“we insist on thinking of having a pain in ocular metaphors – as having a funny sort of particular before the eye of the mind. That particular turns out to be a universal, a quality hypostatized into a subject of predication.”*³⁹⁶ The *“mental-physical distinction”* is based on the *“universal-particular”* distinction and is thus only secondary and parasitic. We consider certain properties (e.g. being red or painful) in isolation and ascribe to them the ability to enter causal relations. In this respect a phenomenal entity doesn’t differ from a Platonic Form.³⁹⁷

³⁹² Ibid., p. 29.

³⁹³ Ibid.

³⁹⁴ Ibid., p. 30.

³⁹⁵ Ibid.

³⁹⁶ Ibid., p. 31.

³⁹⁷ Ibid., pp. 31f.

Thus, we got to a certain illumination of the dualist tendency to “create” new entities (feelings, pains, etc.) that weren’t there in the first place. Wilfrid Sellars, with whom Rorty shares many ideas, comes to a similar conclusion and brings some useful insights in his paper *Empiricism and the philosophy of mind*.³⁹⁸ Sellars suggests that when his mythical Jones develops a theory of sense perception, he first postulates inner theoretical episodes called “*impressions*”, “*which are the end results of impingement of physical objects and processes on various parts of the body, and, in particular, (...) the eye.*” However, it is necessary to stress the following: “*The entities introduced by the theory are states of the perceiving subject, not a class of particulars.*” What might make us think that impressions are introduced as particulars is the fact that a *model* is used in formulation of this “*ur-theory*”: “*This time the model is the idea of a domain of “inner replicas” which, when brought about in standard conditions, share the perceptible characteristics of their physical source. It is important to see that the model is the occurrence “in” perceivers of replicas, not of perceivings of replicas. Thus, the model for an impression of a red triangle is a red and triangular replica, not a seeing of a red and triangular replica.*” When we mistakenly overlook that the model is really only a model, we thereby come to a conclusion that the entities described by it are particulars, just as in the model itself.³⁹⁹

Thanks to reflecting on the way how our conceptual grasp of impressions is built we might come to the conclusion that the concepts in question are “*primarily and essentially intersubjective*”. The fact that we can introspect, have private access to, and report our impressions is only a secondary layer of these concepts, depending on the layer of their intersubjective role – “*...the fact that overt behavior is evidence for these episodes is built into the very logic of these concepts as the fact that observable behavior of gases is evidence for molecular episodes is built into the very logic of molecule talk.*”⁴⁰⁰

As I understand Sellars, his point is that we “discovered” impressions thanks to the fact that we gradually invented “*the language of impressions*”. This language wasn’t formed in order to accurately describe what we knew was already there; quite

³⁹⁸ Sellars, W., *Science, perception and reality*, pp. 127-196.

³⁹⁹ *Ibid.*, p. 191.

⁴⁰⁰ *Ibid.*, p. 195.

the contrary, we started to notice new “dimensions” of our perception thanks to the way we modelled our language of it. Sellars’s mythical Jones “*construes as data the particulars which he has come to be able to observe, and believes them to be antecedent objects of knowledge which have somehow been in the framework from the beginning. It is in the very act of taking that he speaks of the given.*”⁴⁰¹

When we combine Rottry’s and Sellars’s accounts of the roots of our tendency to view conscious sensations as particulars, the following picture pops up: we view states of perceiving as particulars because we use the model of inner replicas. These replicas are, in turn, results of hypostatizing universals abstracted from particular instantiations. I believe that this is an accurate description of how the concept of qualia arises for philosophers, a concept that is apparently very unintuitive for the folk.

In his following analyses, Rorty sees the problem of the distorted view of consciousness as rooted in Cartesian epistemology which is based on the model of intellect *inspecting* “*entities modelled on retinal images*”, whereas in Aristotelian teaching the “*intellect becomes all things*” and is itself thought of as a retinal image. In short, with Descartes, representations come on the scene: “*The Inner Eye surveys these representations hoping to find some mark which will testify to their fidelity.*”⁴⁰² Here arises “*the difference between mind-as-reason and mind-as-consciousness*” – Descartes found a common feature of the mentioned representations, namely “*indubitability*”. The core characteristic of consciousness is “*that there is no distinction between appearance and reality, whereas everywhere else there is.*” Rorty asserts that it is thus possible to conjecture that indubitability explains Descartes’s “*conceptual revolution*”.⁴⁰³ Indubitability started “*to serve as a criterion of the mental*” because pains and thoughts, according to Descartes, share this “*common factor*” and thereby differ from the physical. Thus, they must be “*modes of a single substance*”. Indubitability starts to be the mark “*of something for which the*

⁴⁰¹ Ibid.

⁴⁰² Rorty, R., *Philosophy and the mirror of nature*, p. 45.

⁴⁰³ Ibid., p. 54f.

Greeks had no name – consciousness.”⁴⁰⁴ The Cartesian change is the “*change from mind-as-reason to mind-as-inner-arena*”.⁴⁰⁵

Rorty believes that the mentioned conceptual change is a basis for the above described viewing of sensations and impressions as particulars or entities: “...we would hardly think of a thought or a pain as a thing (a particular distinct from a person, rather than a state of a person) which was not locatable unless we already had the notion of a nonextended substance of which it might be a portion.”⁴⁰⁶ Rorty further notes that the term “substance” might only be an unwise choice of “corrupted scholastic vocabulary” and “that Descartes’s insight was merely a recognition of the difference between parts of persons or states of those parts (e.g., cramps of their stomach) on the one hand and certain states of the whole person on the other”.⁴⁰⁷

Perhaps it would be possible to say (remembering Sellars) that “substance” is only a model which is further elaborated by the introduction of particulars – the replicas which we use in order to conceptually capture impressions. It is only when this model is not recognized as a model and is understood “too literally” that the mind-body problem arises in its whole beauty. For, as Rorty notes, “insofar as dualism reduces to the bare insistence that pains and thoughts have no places, nothing whatever hangs on the distinction between mind and body.”⁴⁰⁸ According to Rorty, there is no metaphysical “problem of consciousness”, but only the epistemological “problem of privileged access”. There is no point in struggle between materialism and dualism.⁴⁰⁹

In order to illustrate his point, Rorty comes with a thought experiment about “the Antipodeans”, beings living far away in the distant part of our galaxy who are like us in almost every respect except for that they don’t know that they have minds. They lack concepts of mental states which would differ from any physical state of a person and never use words like “consciousness”, “mind” or “spirit” to explain the otherwise acknowledged difference between them as persons and other animals.

⁴⁰⁴ Ibid., p. 58.

⁴⁰⁵ Ibid., p. 61.

⁴⁰⁶ Ibid., p. 63.

⁴⁰⁷ Ibid., p. 66.

⁴⁰⁸ Ibid., p. 68.

⁴⁰⁹ Ibid., p. 69.

Instead of reporting feelings and pains they report neural states (e.g., firing of a c-fibre). They don't understand the notion of raw feels and are baffled when asked about them. For example, it is absolutely obvious to them that only the light is of a certain colour, not the neural state they are in when the light reaches their sensory apparatus.⁴¹⁰ Now an important thing to realise is that people from Earth have no way how to determine whether the Antipodeans do or do not have minds. They only speak about stimulated fibres, not about raw feels, pains, or other impressions. Raw feels have something additional – “*a phenomenal property – one which you cannot have the illusion of having (because, so to speak, having the illusion of it is itself to have it)*”, whereas you can have the illusion of having stimulated C-fibres (you can have some different fibres stimulated instead). Thus, the Antipodeans don't go with the appearance-reality distinction in the sense that humans from Earth understand it, since this would require “*a distinction between subjective representations and objective states of affairs*”. The Antipodeans recognize only “*a matter of getting something wrong, having a false belief*”, because no representations have place in their epistemological scheme.⁴¹¹

Humans from Earth (for Rorty, *philosophers* for me) recognize a special type of epistemic situation – it is not possible to misleadingly apprehend a mental phenomenon, while it is possible to misleadingly apprehend a physical phenomenon. We cannot be mistaken about having a pain whereas we can be mistaken about the presence of heat in the outer environment.⁴¹² We isolate a special ontological layer of the mental – “*raw feels, passing thoughts, and mental images*” - which are “*incorrigibly knowable*” to us as their possessors. The crucial question, however, is whether we are really warranted to assert that there are separate ontological categories – the physical and the mental. Are the “*funny extra states besides the neurological ones*” important enough to justify this ontological divide?⁴¹³

The Antipodeans (for Rorty, and to certain extent *the folk* for me) would, indeed, view the “*whole notion of incorrigibly knowable entities, as opposed to being incorrigible about how entities seem to me*” as a matter of twisted and mistaken

⁴¹⁰ Ibid., pp. 70ff.

⁴¹¹ Ibid., p. 77.

⁴¹² Ibid., p. 78.

⁴¹³ Ibid., pp. 80ff.

language. “*Seemings*” as entities of some kind simply don’t exist to them.⁴¹⁴ They do have concepts of states, e.g., of “*feeling*”, but they just don’t have any concept of “*feelings*” – some entities which are to be grasped and become a part of our knowledge. “*Antipodean has the verb but not the noun, so to speak.*”⁴¹⁵ Since the Terrans (philosophers) have both the verb and the noun, they recognize “*a difference between being in a state such that it seems to one that one is (...) and having a raw feel. The former state is an epistemic position toward something about which doubt is possible. The latter state automatically puts one in an epistemic position toward something about which doubt is impossible.*”⁴¹⁶

Rorty attempts to criticize our (philosopher’s) conceptual toolkit by comparing it with that of the Antipodeans. This way he can show how the “*image of the Mirror of Nature – of knowledge as a set of immaterial representations*”⁴¹⁷ distorts our (philosopher’s) view of the mind-body problem and leads us on the wrong track. He simply suggests “*that we abandon the notion that we possess incorrigible knowledge by virtue of a special relation to a special kind of object called “mental objects.” This suggestion is a corollary of Sellars’s attack on the Myth of the Given.*”⁴¹⁸ Thus he doesn’t want to be neither a dualist, sceptic, behaviourist, nor identity-theorist, since all these positions somehow operate with terminology and conceptions that understand “*mind as mirroring nature*” and tend towards Cartesian view that “*the mind is naturally “given” to itself.*” Giving up this image is all we need to do.⁴¹⁹

The folk do not seem to possess this mistaken conceptual toolkit, even though they are generally theory-lite and don’t share the Antipodean approach either. Their view might be somewhere in-between, since they tend towards the naïve view of colours and pains, just as Sytsma suggested.

By dropping the notion of “*special, felt, incommunicable qualities*” by which we come to gain “*direct knowledge*” of our mental states we are equipped to

⁴¹⁴ Ibid., p. 87.

⁴¹⁵ Ibid., p. 92.

⁴¹⁶ Ibid., p. 93.

⁴¹⁷ Ibid.

⁴¹⁸ Ibid., p. 95.

⁴¹⁹ Ibid., p. 97.

understand the problem of privileged access more clearly. It is not the case that we learn what pain is by direct access to the special felt qualities that are subsequently expressed in words as some kind of directly accessed knowledge. Pre-linguistic infant doesn't know about any entities flashing in front of its inner eye; such a child *"knows that it has pain in the way in which the record changer knows the spindle is empty, the plant the direction of the sun, and the amoeba the temperature of the water."*⁴²⁰ Antipodeans would view entities knowledge of which is privileged *and incommunicable* as superfluous and unintelligible. They would not even pass as Wittgensteinian "between somethings and nothings".⁴²¹ As I understand Rorty, he puts forward Antipodean insights because he is convinced, together with Sellars, that sensations are never purely private and never incommunicable – if it wasn't for the public circumstances of learning meanings of words such as "pain", we would never come to an idea of inner representations, particulars, and mental entities, that is, some sort of "somethings".

Another important point to add to the exposition of Rorty's thoughts is that by refuting dualism we don't automatically subscribe to materialistic identity-theory. By claiming that everything mental is physical we unreflectively assume metaphysical position that is not warranted. All that the materialist should be concerned with is the *"predictive or explanatory or descriptive power"* of our concepts. Antipodean language has merit inasmuch it doesn't affect this power. We can afford to drop our superfluous talk of "mental entities" and reduce it to Antipodean without loss of anything important.⁴²²

Refusing the talk of inner mental entities as redundant doesn't entail the victory of "the physical". The point of the whole tossing away of *"the image of man as possessor of a Glassy Essence, suitable for mirroring nature"* is precisely that it doesn't make sense anymore to talk about the two ontological realms. Antipodeans could never *"be able to infer that matter had triumphed over spirit, science over privacy, or anything over anything else. These warring opposites are notions which do not make sense outside of a cluster of images inherited from the Terran*

⁴²⁰ Ibid., p. 110.

⁴²¹ Ibid., pp. 111f.

⁴²² Ibid., p. 120.

seventeenth century.”⁴²³ Indeed, I suggest that the folk, sharing a certain portion of Antipodean spirit, don’t naturally see the problem of consciousness in the same way as philosophers do. Folk dualism dwells in something quite different.

Rorty’s Antipodeans (and my laymen) simply don’t create the distance between the experiencer and her experiences in the way Terran philosophers do. They don’t try to *look at* the subjective conscious states; they simply stay *in them* (the folk in their naïve view, believing that the colour and pains are “out there” in the objects). Conscious awareness is real, and this is not what Antipodeans deny; all they deny or simply refuse to accept is that conscious states are “made of” some *object-like* entities such as raw feels and pains and colourful replicas. No such entities could be *observed*. As soon as we claim that we observe them, we forget about the very act of observation.

As John Searle⁴²⁴ points out, any act of observation is itself a conscious state (and this conscious state inasmuch it is a conscious state simply cannot be observed, in principle).⁴²⁵ The fact that Antipodeans and the folk don’t have the same concepts as philosophers do doesn’t mean that they have no conscious lives; it only shows that they leave consciousness undistorted.

Searle brings forward the true nature of introspection that is often misunderstood. The problem is that we tend to think that the introspection provides us with certain objects (namely “indubitable” inner mental entities) in contrast to perception. However, introspection is not about objects (or particulars), but about focusing our attention on the bare fact that conscious observation is happening.⁴²⁶

Together with Rorty and Sellars we could say that it is more appropriate to speak of states of a person than of some inner particular representations. The image of the inner eye is only a model and taken too literally it becomes inappropriate and misleading. Searle too criticises our tendency to view introspection via ocular metaphors: “*The very fact of subjectivity, which we were trying to observe, makes*

⁴²³ Ibid., pp. 122f.

⁴²⁴ I address Searle’s concept of *ontological subjectivity* in greater depth in my master’s thesis: Košová, M., *Modern Theories of Consciousness and the Elusiveness of Subjectivity*, pp. 32-38. Here I refer to some of his thoughts again.

⁴²⁵ Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge: MIT Press, pp. 96f.

⁴²⁶ Ibid., p. 97.

*such an observation impossible. Why? Because where conscious subjectivity is concerned, there is no distinction between the observation and the thing observed, between the perception and the object perceived. The model of vision works on the presupposition that there is a distinction between the thing seen and the seeing of it. But for “introspection” there is simply no way to make this separation. Any introspection I have of my own conscious state is itself that conscious state.”*⁴²⁷ There is no inner eye that would look at the phenomenal properties of conscious states; the conscious state itself is the looking.

We cannot observe our conscious states because we, as subjects, are “in them” (in the “zero distance” from them).⁴²⁸ This resonates nicely with Rorty’s criticism of the idea of “glassy essence” and “the mirror of nature”. His Antipodeans (and the folk) stay in the “zero distance” from their mental states and thus do justice to their ontological status. Antipodeans and the folk are in their mental states and don’t attempt to step outside of themselves.⁴²⁹ Terran philosophers, on the other hand, attempt to “escape their own shadow”, so to speak.

Our concept of observation includes the idea of observers (*objective in the epistemic sense*) who observe the *ontologically objective* reality. Subjectivity present in the *act of observation* transcends the observed outer reality. We simply cannot observe this act itself, because it provides us with “*the subjective (ontological sense) access to objective reality.*” It is possible for me to observe persons around me, but I will never be able to observe their (or even my own) *subjectivity*: “*The ontology of observation – as opposed to its epistemology – is precisely the ontology of subjectivity.*”⁴³⁰ When the folk encounter colours and pains, they seem to view the situation as a regular case of observation of the outer reality (the naïve view proposed by Sytsma). They don’t postulate another layer of “reality” to observe, and thus avoid the talk of qualia, etc.

⁴²⁷ Ibid.

⁴²⁸ Košová, M., *Modern Theories of Consciousness and the Elusiveness of Subjectivity*, p. 33.

⁴²⁹ “*Whereof one cannot speak, thereof one must be silent.*” (Wittgenstein, L. (1960). *Tractatus logico-philosophicus*. (C. K. Ogden, Trans.) London: Routledge & Kegan Paul Ltd., p. 189.)

⁴³⁰ Searle, J. R., *The rediscovery of the mind*, pp. 98f.

This might be the reason why the concept of conscious will and the connection of consciousness to the concept of free will seems to be so obvious and yet so rarely explained. The folk know they are subjects of conscious experience, but they don't have a proper concept of the extra layer of the phenomenal properties. Ontological subjectivity of the conscious awareness provides the self with the transcendent unity, but there is no layer of phenomenal states or qualia which would create an explanatory gap for the folk.

The folk concept of a person is not dependent on conscious states as such and their irreducibility (as in the case of philosophical qualia), but certain specific type of conscious states that qualify the being in question as worth of moral considerations. The ability to feel Strawsonian reactive emotions, ability to feel pain, and caring about how things turn out for me are the examples of conscious states that play role in folk dualism. What is more, the folk concept of free will seems to go hand in hand with this type of conscious states.

Conclusion

More research is needed to explore further details about relationships between the concepts of *free causal agent*, *essential moral self*, and *conscious subject*, but I have attempted to show all these aspects of the concept of self as interwoven building blocks of the folk concept of a person.

Firstly, I have introduced the concept of self from a wider perspective and confronted different philosophical views in order to show the necessity to bring the idea of different frameworks into the picture. We don't need to invoke McGinn's transcendental naturalism, nor do we have to claim that the self is nothing more than a mere illusion. When we realise that our "dualism" stems from the collision of the two different frameworks we use to approach the world, we can acknowledge them both as equally important and valuable.

Secondly, I have addressed the main aspects of the concept of a person one by one to provide a more-detailed notion and to demonstrate how experimental approach benefits the attempts to capture the concept properly. While addressing the problem of free will, we revealed dualistically coloured concept of *autonomous agent* that escapes causal laws of scientific framework. In order to do justice to the folk concept of free will we had to acknowledge that it has certain "supernatural" characteristics. I also suggested that the basis for this folk notion might be the idea that any free act of a human agent has to be somehow connected to conscious mental states (*conscious will* that provides the ultimate unity of the self). Experimental philosophy supports the view that the folk notion of free will doesn't fit the scientific framework and that people connect free action with the vocabulary of the Davidsonian mental realm, even though the folk are theory-lite with regard to the details of this specific "transcendence" of autonomous agents. Connection of this notion with consciousness also gained support and clarification from x-phi (later in the fourth chapter).

In the third chapter I turned to the concept of *essential self* and explored the folk concept of a person from the perspective of personal identity and its connection to the folk concept of soul. I have shown that the concept of the *essential moral self* (that comes forward as a clear result of experimental studies of personal identity across many different cultures) shares crucial characteristics with the folk concept of

soul. These crucial characteristics are connected to moral traits that play role in interpersonal relationships. The concept of mind proved to differ from the concept of soul to a large extent: while the soul is viewed as more independent from the physical world, more connected to personal identity and to moral dimension, the mind is usually associated with cognitive abilities that are morally neutral. It shares certain characteristics with the soul but also with the view people have about the physical brain. The implicit dualist tendencies reveal themselves in the context of contemplating the function of the brain: precisely those traits that characterize the concept of essential moral self and soul are ascribed to the brain to a significantly lesser extent than morally neutral cognitive and physical abilities. Even though people generally seem to accept scientific description of persons, they are reluctant to fully acknowledge that also the essential moral self falls under this description. I have once again demonstrated that the attempt to reduce the manifest image to the scientific image fails when we stand face to face with moral and interpersonal dimension of human existence.

Finally, in the fourth chapter I have addressed the last sub-concept of the folk concept of self – the *conscious subject*. I reviewed a theory that claims that the phenomenal consciousness (as it is traditionally understood by philosophers of mind) plays a role in folk dualism. With help of a battery of experimental philosophy studies I have attempted to show that the situation is much more complex: research in the area of the folk concept of consciousness suggests that lay people don't have proper concept of qualia. Rather, they seem to lean towards the naïve view of colours and pains (they tend to believe that pains and colours are situated in the outside objects, not inside the mind). What is more, it is not the phenomenality of the conscious states and the irreducibility of qualia that creates the explanatory gap within the context of folk dualism, but a specific type of conscious states: moral emotions, ability to feel pain or pleasure, and more generally the fact that the being cares for how things turn out for her are the right kinds of conscious states that adorn their owner with the status of a person worth of moral considerations, a being that is not graspable by the language of science.

We have seen that folk dualism arises in front of us as a tendency to understand persons as escaping physically described world within its own transcendent realm of mental states revolving around morality of interpersonal relationships. Even though cognitive abilities and other kinds of traits play

undeniable role in the identity of persons, moral and normative interpersonal dimension showed as the central motif that connects all three aspects of the concept of self: doing justice to our *essential selves* (or souls) means to nurture traits that have positive effects within and that are approved by our human community. Our status of a morally responsible *free agent* and a member of community depends on the fact that we are capable of experiencing *conscious states* connected to interpersonal relationships (Strawsonian reactive emotions) and that we care for how things turn out for us and for other sentient beings around us (able to feel pain and pleasure of all sorts).

Together with the progress of science we often encounter attempts to explain a whole range of human behaviour through scientific framework while we view this framework as the only “true” one. A human being is thus “nothing more” than interplay of molecules and neuronal firings and a mere animal lost in the illusion of moral supremacy. In the face of fascinating scientific studies, we come to “realise” that free will is not possible and the world of responsible moral agents is just a complex fabrication of our brains. But is this the only truth there is?

Together with Davidson and Sellars (and Kant before them) I have attempted to defend the view that it is not possible to fully grasp a human being within the boundaries of the scientific image and the physical realm. We need to keep another framework in order to understand what it means to be a person. This framework has its own logic and rules which are tightly connected to the fact that the folk (non-philosophers – the majority of people in the world living their everyday lives) are theory-lite with regard to their worldview. The picture they need in order to lead their human lives works just fine when it is blurry.⁴³¹ Thus people believe that a human agent together with her beliefs, reasons, desires, and emotions connected to what matters to her escapes physical laws of causation, without providing precise physical description of how this could possibly work.

The framework that captures the concept of self in its due depth is the source of folk dualism in the moment when it is confronted with the scientific framework and thus forced to defend itself at all costs. When we fully accept strict physical laws and the model of causation that leaves no space for the language of the mental realm,

⁴³¹ Once again, a reference to Wittgenstein, L., *Philosophical investigations*, p. 34 (§71).

autonomous moral agents start to flow above the causal chain as ghosts from fantastic stories. When we fully accept that the true identity of a human being dwells in the way her brain and body function, human soul becomes an immaterial entity from the “other world”. And when we define the deepest human moral emotions and pains as a specific brain activity, morality becomes some kind of “divine sparkle” that transcends this physical realm. These aspects of human beings are no longer natural, and they are forced on the very outskirts or even behind the “boundaries” of the “real” world.

Inspired by the approach of Davidson and Sellars I have proposed that the key to avoiding this abyss between man and the physical world is to leave the two frameworks exist in parallel and don’t try to reduce either of them to the other. We need to realize that each of the frameworks serves us for different purposes. Science helps us to reveal hidden mechanisms behind natural phenomena, and it provides us with priceless tools for improving the quality of life. However, we should not talk about the physical causes within human brain when describing actions of a person in everyday situations. Instead, we should talk about reasons, beliefs, motivations, and emotions. When we talk about the feeling a person has towards her close ones, we should not talk about the brain activity, but about the person’s deep essential self. After all, most of the time we cannot help exposing our implicit dualist beliefs.⁴³² If we want to remain free moral beings, we need to save the framework that provides the right atmosphere for persons and acknowledge its own truth.

The importance of saving the manifest image, the mental realm, and folk dualism (all three being slightly different aspects of one conceptual framework) shows as inevitable not only for preserving what we value most about human beings for today, but also for the future, face to face with the ascend of future technologies. The problem of personal identity is often addressed in the literature on future technologies, but the folk concept revealed by experimental approach (the importance of moral and interpersonal traits) remains deeply neglected.⁴³³ I believe

⁴³² See e.g. Mudrik, L., & Maoz, U. (2015). “Me and my brain”: Exposing neuroscience’s closet dualism. *Journal of Cognitive Neuroscience*, 27(2), 211-221.

⁴³³ See e.g. Schneider, S. (2019). *Artificial you*. Princeton: Princeton University Press.

that folk concepts are crucial for addressing these questions. As I have already suggested in the introduction, the reason why I find folk intuitions important is the fact that I see value in philosophy attempting to keep in touch with the reality of everyday human lives and real-life moral challenges. Philosophy (especially philosophy addressing moral issues) should serve people in identifying what is important for human beings and how they should act with regard to humanity and all sentient beings in general.

As an example, let's briefly consider the problem of human enhancement. The main goal of transhumanism⁴³⁴ is to use technological progress to improve the quality of human life, especially by removing suffering: "*Advances in genetic engineering, artificial intelligence, robotics and nanotechnology (...) allow us to conquer disease, eliminate unhappiness, end scarcity and postpone, perhaps indefinitely, death itself.*"⁴³⁵ There are many possibilities how to enhance a human being: we can use new technologies to improve intelligence, physical strength, physical appearance, physical and mental health, but also the ability to experience certain emotions or certain predispositions concerning our conduct in morally-laden situations. An important question is what traits it is safe to change in the face of what we know about the folk concept of a person. It would be probably permissible to enhance physical and cognitive traits to some extent, but is it morally right to change

I had an opportunity to talk to Susan Schneider in person when she was a keynote speaker at the Ernst Mach Workshop in 2019. I asked her about her opinion on the role of the folk concept of personal identity in the debates about the possible impact of future technologies. She was not familiar with the studies that explore folk concept of personal identity, and when I mentioned experimental philosophy, she added that she doesn't think that folk intuitions could be of any help in solving these questions.

⁴³⁴ For more information about the movement visit the websites of the World Transhumanist Association: www.transhumanism.org, or the Extropy Institute: www.extropy.org (websites recommended by Ch. T. Rubin as representing the most influential transhumanist organisations).

⁴³⁵ Rubin, C. T. (2008). What is the good of transhumanism? In B. Gordijn & R. Chadwick (Eds.), *Medical Enhancement and Transhumanity* (pp. 137-156). Springer Science + Business Media B. V., p. 137.

our deep moral true selves?⁴³⁶ Or isn't it morally desirable, or even a moral imperative to enhance our moral selves and thus come closer to the true-self concept? Shouldn't we strive to be "humane" (cultivate compassion, empathy, and strive to be better people) instead of just being "human" - "normal" and "natural" with all the bad things that come with our natural state of being?⁴³⁷

On the other hand, just as Rubin suggests, the possibility of transhumanism could pose a threat to many traditional views and values, "*any that depend on the existence of a soul, for example, or divine revelation or on a given human nature, or even on existent social constellations.*"⁴³⁸ In reply to Bostrom he presents a very important point: we are capable of being humane, of feeling compassion and empathy precisely because we are not perfect. We have mortal bodies and minds susceptible to suffering of all sorts, and this enables us to reach a deep understanding of what it means to suffer and to empathise with other suffering sentient beings. What is more, we value everyone who strives to be a better human being precisely because "*it is hard to be better*". Do we really need upgrades to help us fight against our dark side, or do we want to be the free agents who do justice to their true selves through their own will?⁴³⁹

I believe that we cannot solve similar questions via purely scientific description of man. Without reflecting upon folk intuitions and without turning to real life and real persons we lose what is really human and humane. The answer to what makes us transcend the world described by science dwells in everyday real-life

⁴³⁶ See e.g. Riis, J., Simmons, J. P., & Goodwin, G. P. (2008). Preferences for psychological enhancements: The reluctance to enhance fundamental traits. *Journal of Consumer Research*, 35(3), 495–508; and Wagner, K., Maslen, H., Oakley, J., & Savulescu, J. (2018). Would you be willing to zap your child's brain? Public perspectives on parental responsibilities and the ethics of enhancing children with transcranial direct current stimulation. *AJOB Empirical Bioethics*, 9(1), 29-38.

Both studies suggest that people express stronger tendency to refuse enhancement of "fundamental" traits, such as kindness and empathy, in contrast to more neutral traits, such as mathematical ability.

⁴³⁷ Bostrom, N. (2003). *The transhumanist FAQ: A general introduction* (Version 2.1), p. 36. Retrieved from <https://www.nickbostrom.com/views/transhumanist.pdf>

⁴³⁸ Rubin, C. T., *What is the good of transhumanism?*, p. 143.

⁴³⁹ *Ibid.*, p. 148.

practices within human communities and in concepts that enable functioning of human societies. I have attempted to show that moral traits playing role in interpersonal relationships, ability to experience emotions connected to moral and interpersonal situations, being able to care how things turn out for me and other sentient beings, and ability to act freely within the logic of mental realm are central to understanding of what people value most about human beings.

This doesn't mean at all that other aspects of humans are less important - different cognitive and physical abilities are no doubt a necessary condition for persons to be able to act within human communities *as persons*.⁴⁴⁰ My aim here, however, was to point to those aspects of persons that remain intuitively inaccessible to scientific explanations (people are happy to ascribe rational thinking to the brain, while the essential moral self traits seem to pose a problem in this context). I attempted to show what seems to constitute "souls" of persons, without claiming that there actually are such entities somewhere "in" or "outside" this world. They exist in the context of everyday human life, a context that philosophy should not, in my view, overlook.

⁴⁴⁰ For example, Daniel Dennett suggests six "themes" that are connected to the search for "*a necessary condition for personhood*": persons are 1. *rational beings*, 2. beings to whom we ascribe mental or *intentional predicates*, 3. beings toward which a certain *attitude or stance* is adopted (by other persons), 4. beings able of *reciprocating* the adoption of an appropriate stance, 5. beings capable of *verbal communication*, and 6. beings that are *conscious* in a certain characteristic way (self-consciousness). (Dennett, D. C., *Brainstorms*, pp. 269f.)

Bibliography

Alexander, J., Mallon, R., & Weinberg, J. M. (2014). Accentuate the negative. In J. Knobe & S. Nichols (Eds.), *Experimental philosophy (Volume 2)* (pp. 31-50). New York: Oxford University Press.

Anglin, S. M. (2014). I think, therefore I am? Examining conceptions of the self, soul, and mind. *Consciousness and Cognition*, 29, 105-116.

Arico, A. (2010). Folk psychology, consciousness, and context effects. *Review of Philosophy and Psychology*, 1(3), 371-393.

Arico, A., Fiala, B., Goldberg, R. F., & Nichols, S. (2011). The folk psychology of consciousness. *Mind and Language*, 26(3), 327-352.

Bartneck, C., & Hu, J. (2008). Exploring the abuse of robots. *Interaction Studies*, 9(3), 415-433.

Bartneck, C., Van Der Hoek, M., Mubin, O., & Al Mahmud, A. (2007, March). "Daisy, daisy, give me your answer do!" Switching off a robot. *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction, Washington DC* (pp. 217-222). New York, NY: ACM.

Baumeister, R., Masicampo, E. J. C, DeWall, C. N. (2009). Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness. *Personality and Social Psychology Bulletin*, 35(2), 260-268.

Bering, J. M. (2006). The folk psychology of souls. *Behavioral and Brain Sciences*, 29(5), 453-462; discussion 462-498.

Bering, J., & Bjorklund, D. (2004). The natural emergence of reasoning about the afterlife as a developmental regularity. *Developmental Psychology*, 40(2), 217-233.

Berniūnas, R. (2012). Folk concept of 'a person': Structure and warrant. *Problemos*, (Supplementary), 63-77.

Blackburn, S. (2005). *The Oxford dictionary of philosophy* (2nd ed.). Oxford: Oxford University Press.

Bloom, P. (2004). *Descartes' baby: How the science of child development explains what makes us human* [Adobe Digital Editions version]. ISBN 9781446473627.

Bostrom, N. (2003). *The transhumanist FAQ: A general introduction* (Version 2.1), p. 36. Retrieved from <https://www.nickbostrom.com/views/transhumanist.pdf>

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Chalmers, D. J. (2010). *The character of consciousness*. New York: Oxford University Press.

Chalmers, D. J. (2018). The meta-problem of consciousness. *Journal of Consciousness Studies*, 25(9-10), 6-61.

Chlup, R. (Ed.). (2007). *Pojetí duše v náboženských tradicích světa* [The conception of soul in the religious traditions of the world]. Prague: DharmaGaia.

Chomsky, N. (2009). The mysteries of nature: How deeply hidden? *The Journal of Philosophy*, 106(4), 167-200.

Curry, O. S., Alfano, M., Brandt, M. J., & Pelican, C. (2020, June 9). Moral molecules: Morality as a combinatorial system. Retrieved from <https://doi.org/10.31219/osf.io/xnstk> (preprint version 4).

Davidson, D. (2001). *Essays on actions and events* (2nd ed.). New York: Oxford University Press.

De Freitas, J., & Cikara, M. (2018). Deep down my enemy is good: Thinking about the true self reduces intergroup bias. *Journal of Experimental Social Psychology*, 74, 307-316.

De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2018). Moral goodness is the essence of personal identity. *Trends in Cognitive Sciences*, 22(9), 739-740.

De Freitas, J., Sarkissian, H., Newman, G. E., Grossman, I., De Brigard, F., Luco, A., Knobe, J. (2018). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science*, 42(S1), 134-160.

De Freitas, J., Tobia, K., Newman, J. E., & Knobe, J. (2017). Normative judgements and individual essence. *Cognitive Science*, 41(S3), 382-402.

Dennett, D. C. (1981). *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: The MIT Press.

Dennett, D. C. (1984). *Elbow room: the varieties of free will, worth, wanting*. London: Clarendon Press.

Dennett, D. C. (1991). *Consciousness explained*. Boston: Back Bay Books.

Dennett, D. C. (2003). *Freedom evolves*. New York: Viking.

Díaz, R. (in press). Do people think consciousness poses a hard problem? Empirical evidence on the meta-problem of consciousness. *Journal of Consciousness Studies*. Advance online publication retrieved from <https://philarchive.org/archive/DAZDPT>

Dranseika, V. (2017). On the ambiguity of 'the same person'. *AJOB Neuroscience*, 8(3), 184-186.

Estes, D. (2006). Evidence for early dualism and more direct path to afterlife beliefs. *Behavioral and Brain Sciences*, 29(5), 470.

Fiala, B., Arico, A., & Nichols, S. (2011). On the psychological origins of dualism: Dual-process cognition and the explanatory gap. In E. Slingerland & M. Collard (Eds.), *Creating consilience: Integrating the sciences and the humanities*. New York: Oxford University Press. Online version retrieved from http://www.u.arizona.edu/~arico/PsychOriginsDualism_final_.pdf

Fiala, B., Arico, A., & Nichols, S. (2014). You Robot. In E. Machery & E. O'Neill, (Eds.), *Current controversies in experimental philosophy* (pp. 31-47). Abingdon: Routledge.

Furley, D. J. (1956). The early history of the concept of soul. *Bulletin of the Institute of Classical Studies*, 3, 1-18.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Sciences*, 4(1), 14-21.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. New York: Oxford University Press.

Gregg, V. R., Winer, G. A., Cottrell, J. E., Hedman, K. E., & Fournier, J. S. (2001). The persistence of a misconception about vision after educational interventions. *Psychonomic Bulletin & Review*, 8(3), 622-626.

Harris, S. (2012). *Free will*. New York: Free Press.

Haslam, N., Kashima, Y., Loughnan, S., Shi, J., & Suitner, C. (2008). Subhuman, inhuman, and superhuman: contrasting humans and nonhumans in three cultures. *Social Cognition*, 26(2), 248–258.

Heiphetz, L., Strohminger, N., Gelman, S., & Young, L. (2018). Who am I? The role of moral beliefs in children's and adults' understanding of identity. *Journal of Experimental Social Psychology*, 78, 210-219.

Heiphetz, L., Strohminger, N., & Young, L. L. (2017). The role of moral beliefs, memories, and preferences in representations of reality. *Cognitive Science*, 41(3), 744-767.

Heubner, B. (2010). Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? *Phenomenology and the Cognitive Sciences*, 9(1), 133–155.

Ichikawa, J. J. (2016). Intuitive evidence and experimental philosophy. In J. Nado (Ed.), *Advances in experimental philosophy and philosophical methodology* (pp. 155-173). Bloomsbury. Online version available from <https://philarchive.org/archive/ICHIEAv1>

Jirout Košová, M. (2020). Skúmanie významu experimentálnej filozofie skrze koncept osobnej identity [Exploring the significance of experimental philosophy through the concept of personal identity]. *Filosofický časopis*, 68(4), 581-603.

Jirout Košová, M., Kopecký, R., Oulovský, P., Nekvinda, M., & Flegr, J. (in press). My friend's true self: Children's concept of personal identity. *Philosophical Psychology*. Advance online publication (preprint version 3) retrieved from psyarxiv.com/uwa59

Johnson, S., Slaughter, V., & Carey, S. (1998). Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. *Developmental Science*, 1(2), 233-238.

Kant, I. (2004). *Fundamental principles of the metaphysic of morals* [EBook #5682]. (T. K. Abbott, Trans.). Retrieved from <https://www.gutenberg.org/files/5682/5682-h/5682-h.htm>

Kauppinen, A. (2014). The rise and fall of experimental philosophy. In J. Knobe & S. Nichols (Eds.), *Experimental philosophy (volume 2)* (pp. 3-29). New York: Oxford University Press.

Knobe, J. (2014). *Free will and the scientific vision*. In E. Machery, E. O'Neill (Eds.), *Current controversies in experimental philosophy* (pp. 69-85). Abingdon: Routledge.

Knobe, J., & Nichols, S. (2008). An experimental philosophy manifesto. In J. Knobe & S. Nichols (Eds.), *Experimental philosophy* (pp. 3-14). New York: Oxford University Press.

Knobe, J., & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, 7(1), 67-83. Advance online publication retrieved from <https://cpb-us-w2.wpmucdn.com/campuspress.yale.edu/dist/3/1454/files/2016/02/Consciousness-28f03o2.pdf>

Košová, M. (2014). *Modern theories of consciousness and the elusiveness of subjectivity* (Master's thesis). Available from <https://is.cuni.cz/webapps/zzp/detail/136799/>

Košová, M. (2014). Skutočná podstata ja [The true nature of the self]. *Pro-Fil*, (Supplementary), 50–64. Available from <http://www.phil.muni.cz/journals/index.php/profil/article/view/998>

Košová, M. (2015). Compatibilism and conscious will. *Filosofie dnes*, 7(1), 61-75. Available from <https://filosofiednes.ff.uhk.cz/index.php/hen/issue/view/15>

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Locke, J. (2009). *Essay Concerning Human Understanding*. WLC books. (Originally published in 1690)

McGinn, C. (1993). *Problems in philosophy: The limits of inquiry*. Oxford: Blackwell.

Mudrik, L., & Maoz, U. (2015). “Me and my brain”: Exposing neuroscience’s closet dualism. *Journal of Cognitive Neuroscience*, 27(2), 211-221.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83(4), 435-450.

Nahmias, E., Allen, C. H., & Loveall, B. (2020). When do robots have free will? Exploring the relationships between (attributions of) consciousness and free will. In B. Feltz, M. Missal & A. Sims, (Eds.), *Free will, causality, and neuroscience* (pp. 57-80). Brill. Retrieved from <https://brill.com/view/title/38676>

Nahmias, E., & Murray, D. (2010). Experimental philosophy on free will: An error theory for incompatibilist intuitions. In J. Aguilar, A. Buckareff & K. Frankish (Eds.), *New waves in philosophy of action* (pp. 189-216). Hampshire, England: Palgrave-Macmillan.

Nahmias, E., & Thompson, M. (2014). A naturalistic vision of free will. In E. Machery & E. O’Neill (Eds.), *Current controversies in experimental philosophy* (pp. 86-103). Abingdon: Routledge.

Nichols, S., & Bruno, M. (2010). Intuitions about personal identity: An empirical study. *Philosophical Psychology*, 23(3), 293-312.

Peregrin, J. (2018). Davidson and Sellars on “two images”. *Philosophia*, 46(1), 183-192.

Peressini, A. (2013). Blurring two conceptions of subjective experience: Folk versus philosophical phenomenality. *Philosophical Psychology*. Advance online publication. DOI: 10.1080/09515089.2013.793150.

Phelan, M., Arico, A., & Nichols, S. (2013). Thinking things and feeling things: on an alleged discontinuity in the folk metaphysics of mind. *Phenomenology and the Cognitive Sciences*, 12(4), 703-725.

Prinz, J., & Nichols, S. (2016). Diachronic identity and the moral self. In J. Kiverstein (Ed.), *The Routledge handbook of philosophy of the social mind* (pp. 449-463). Abingdon: Routledge.

Richert, R. A., & Harris, P. L. (2006). The ghost in my body: Children's developing concept of the soul. *Journal of Cognition and Culture*, 6(3-4), 409-427.

Richert, R. A., & Harris, P. L. (2008). Dualism revisited: Body vs. mind vs. soul. *Journal of Cognition and Culture*, 8(1-2), 99-115.

Riis, J., Simmons, J. P., & Goodwin, G. P. (2008). Preferences for psychological enhancements: The reluctance to enhance fundamental traits. *Journal of Consumer Research*, 35(3), 495–508.

Rorty, R. M. (1979). *Philosophy and the mirror of nature*. Princeton: Princeton University Press.

Rose, D., & Nichols, S. (2019). Teleological essentialism. *Cognitive Science*, 43(4), e12725. doi:10.1111/cogs.12725

Rubin, C. T. (2008). What is the good of transhumanism? In B. Gordijn & R. Chadwick (Eds.), *Medical Enhancement and Transhumanity* (pp. 137-156). Springer Science + Business Media B. V.

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., Sirker, S. (2010). Is belief in free will a cultural universal? *Mind & Language*, 25(3), 346-358.

Schneider, S. (2019). *Artificial you*. Princeton: Princeton University Press.

Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge: MIT Press.

Sellars, W. (1963). *Science, perception and reality*. London: Routledge & Kegan Paul Ltd.

Shepherd, J. (2015). Consciousness, free will, and moral responsibility: Taking the folk seriously. *Philosophical Psychology*, 28(7), 929-946.

Smart, J. J. C. (1959). Sensations and brain processes. *The Philosophical Review*, 68(2), Ithaca, NY: Cornell University, 141-156.

Starmans, C., & Bloom, P. (2018). Nothing personal: what psychologists get wrong about identity. *Trends in Cognitive Sciences*, 22(7), 566-568.

Strawson, F. P. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48, 1-25.

Strohming, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*, 12(4), 551-560.

Strohming, N. & Nichols, S. (2014). The essential moral self. *Cognition*, 131(1), 159-171.

Strohming, N., & Nichols, S. (2015). Neurodegeneration and identity. *Psychological Science*, 26(9), 1469 - 1479.

Sytsma, J. (2010). Dennett's theory of the folk theory of consciousness. *Journal of Consciousness Studies*, 17(3-4), 107-130. Advance online version retrieved from http://philsci-archive.pitt.edu/5141/1/Dennett%27s_Theory_of_the_Folk_Theory_of_Consciousness.pdf

Sytsma, J. (2014). Attributions of consciousness. *WIREs Cognitive Science*, 5(6), 635-648.

Sytsma, J. (2014). The robots of the dawn of experimental philosophy of mind. In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy* (pp. 48-64). Abingdon: Routledge.

Sytsma, J., & Machery, E. (2010). Two conceptions of subjective experience. *Philosophical Studies*, 151(2), 299-327.

Tobia, K.P. (2015). Personal identity and the Phineas Gage effect. *Analysis*, 75(3), 396-405.

Tobia, K.P. (2016). Personal identity, direction of change, and neuroethics. *Neuroethics*, 9(1), 37-43.

Wagner, K., Maslen, H., Oakley, J., & Savulescu, J. (2018). Would you be willing to zap your child's brain? Public perspectives on parental responsibilities and the ethics of enhancing children with transcranial direct current stimulation. *AJOB Empirical Bioethics*, 9(1), 29-38.

Walter, S. (2014). Willusionism, epiphenomenalism, and the feeling of conscious will. *Synthese*, 191(10), 2215-2238.

Wegner, D. M. (2005). Who is the controller of controlled processes? In R. R. Hassin, J. S. Uleman & J. A. Bargh (Eds.), *The new unconscious: Social cognition and social neuroscience* (pp. 19-36). New York: Oxford University Press.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental live. *Proceedings of the National Academy of Sciences of the United States of America*, 114(43), 11374-11379.

Wittgenstein, L. (1958). *Philosophical investigations* (2nd ed.). Oxford: Basil Blackwell.

Wittgenstein, L. (1960). *Tractatus logico-philosophicus*. (C. K. Ogden, Trans.) London: Routledge & Kegan Paul Ltd.